

A MODEL FOR LARGE SPARSE CONTINGENCY TABLES*

Joseph Charles Marhoul

Department of Statistics

and

Stanford Linear Accelerator Center

Stanford, California

LCS Technical Report No. 13

December 1984

ABSTRACT

The intent is to indicate that non-linear regression with categorical predictors yields a reasonable methodology for analyzing large sparse tables.

* Work supported by a NSF Mathematical Sciences Postdoctoral Research Fellowship, Office of Naval Research contract N00014-83-K-0472 and grant N00014-83-G-0121, and U.S. Army Research Office under contract DAAG29-82-K-0056

Chapter 1

Introduction

Perhaps one of the main impediments to rapid progress in the development of the social, behavioral and biological sciences is the omnipresence of qualitative data. All too often it is simply impossible to obtain numerical data; the researcher has the choice of qualitative data or no data at all. [31]

This apology (or excuse) indicates the need of analyzing categorical data. A specific case in which categorical data makes an appearance is a contingency table. Although the literature is replete with methodologies for the analysis of small size contingency tables, relatively few methods have been proposed for the case when there are many cells of the tables and many missing or zero values. The intent of this thesis is to indicate that non-linear regression with categorical predictors yields a reasonable methodology for analyzing such large sparse tables.

The first section of this chapter introduces the concepts of categorical variables and of contingency tables. It is included purely for the sake of completeness and is intended to neither enlighten nor stimulate the knowledgeable reader. The second section offers some history of the methodology used to study contingency tables, and indicates possible shortcomings of these procedures when applied to large sparse tables.

1.1. Basics.

A categorical variable is a measurable mapping from some measure space to a finite

set. The elements of the range are called categories. Classifications such as sex, occupation, race, and marital status immediately present themselves as examples of categorical variables. Male, female, sometimes, and never are examples of specific categories of the classification "sex". Any random variable can be transformed into a categorical variable. Let X be a random variable, $\{P_j\}_{j=1}^n$ a partition of the range of X , and $\{\alpha_j\}_{j=1}^n$ a set. Then the mapping

$$C(X) = \alpha_i \quad \text{if } X \in P_i \quad (1.1.1)$$

is categorical.

From this construct it is clear that all data may be considered categorical, since any measurement is inherently of finite precision and consequently can only be reported as lying in a certain interval. Although this view may seem a bit extreme and counter-productive, it is not without its proponents [31]. The approach taken here is to consider any mapping derived by (1.1.1) to be categorical only if the number of categories is "reasonably small". What constitutes "reasonably small" is left undefined, it being assumed the reader is sufficiently intelligent to interpret it in a "reasonable" manner.

Let S be an arbitrary set, and $\{C_j\}_{j=1}^n$ be mappings $C_k: S \rightarrow \Gamma_{m_k}$, where Γ_{m_k} is some set of m_k distinct symbols. The relation \sim may be defined on S by $s_1 \sim s_2 \Leftrightarrow C_j(s_1) = C_j(s_2) \forall j$. It is easy to show that \sim is an equivalence relationship. A $m_1 \times m_2 \times \dots \times m_n$ contingency table is defined to be any function g which is constant on these equivalent classes. It is noted that the functions $\{C_j\}$ are categorical variables. Consequently a contingency table may also be thought of as a function in which all points having the same categorical predictors are mapped into a single value. Such tables arise frequently from the cross-classification of a population according to several characteristics. In this instance the set S would be the set of individuals in the population, with C_k corresponding to the k^{th} characteristic of the cross-classification, $\{\Gamma_i\}_{i=1}^{m_k}$ the distinct instances of the k^{th} characteristic, and $g(\cdot)$ being either the number of individuals in the population having the specified traits or the proportion of individuals in the population having the specified traits. The elements of the range in the former case are known as counts and in the latter proportions.

It is possible to consider this a particular instance of a regression model where the existence of a mapping $f: S \rightarrow \mathbb{R}$ is assumed and the random variables observed are the $(n + 1)$ -tuples $(C_1(s), \dots, C_n(s), f(s))$. This differs from the contingency table in that it is not assumed that $C_j(s_1) = C_j(s_2) \forall j$ implies $f(s_1) = f(s_2)$. Thus, points having the same categorical predictors are allowed to have different responses. This type of data commonly arises when a population is cross-classified according to a set of characteristics after which some type of experiment is performed and a continuous response is observed for each member. It should be noted that in this set-up a contingency table still may be appropriate if the response is discrete, the particular response may be considered another classification of the population.

Given a $m_1 \times m_2 \times \dots \times m_n$ contingency table G it is possible to construct a $m_1 \times m_2 \times \dots \times m_n$ table whose $(i_1, i_2, \dots, i_n)^{th}$ entry is $G(i_1, i_2, \dots, i_n)$. Any entry not in the range of G is called a missing value and any entry which is zero will be called a zero entry. These entries arise from several different causes. Two of the most common ones are:

1. Because of the finiteness of the sample size an entry which would not be zero if an infinite sample size was drawn is zero. These are known as zeros due to sampling variation.
2. Due to the classification of the entries certain combinations may be impossible or redundant. Such zeroes are known as structural zeroes.

Zeroes due to sampling variation occur frequently when the sample size and the total number of elements in the table are of the same order of magnitude, and is a common occurrence when the population is classified according to many characteristics. A nationwide educational survey may not have anyone classified as a white, eastern European Jewish male between twenty-five and thirty years of age who farms in the midwest, although there do exist people belonging to this category.

A natural example of structural zeroes may be found in genetics. Certain allele combinations are known to be fatal and thus classifying animals according to these genotypes will

necessarily yield zeroes counts for certain combinations. These entries are known *a priori* to be zero and any analysis of the data should take this into account.

There have been many methods of analyzing contingency tables which have no zero cells, and many of these methods have been successfully adopted to the case of structural zeroes. The same is not quite true for the case of sampling zeroes, and it is this case which shall be of primary interest in this work.

1.2. Brief Overview.

The literature of contingency tables is vast and scattered. No attempt has been made to be thorough and the topics presented are more likely to represent the author's preference rather than any concept of "importance".

The use of contingency tables dates back at least to the early nineteenth century with the work of Quetelet [26]. The actual analysis of contingency tables, however, is considered to have begun with Pearson [25], who first proposed the classical χ^2 test. Pearson adopted the view that categorical variables could always be thought of as a discretization of some possible unknown continuous random variable and insisted that all analysis be based on this assumption. This necessitated some logical acrobatics to deal with apparently dichotomous predictors. Hence Pearson considered *live* vs. *dead* to be extreme values of some continuous scale of "health" and argued that *employed* vs. *unemployed* were a discretization of some continuous "amount of work" variable.

In the same year Yule [32] proposed another theory for contingency tables. His view was that the categorical variables were fixed and did not arise from any discretization process. This was in direct opposition to Pearson's work, leading to a long and bitter correspondence between the two. For a long period of time during and following this debate the study of contingency tables was constrained to 2×2 tables. Progress was made in 1935 when Bartlett [3] derived a definition for second order interactions in $2 \times 2 \times 2$ tables .

To facilitate the discussion, the following notation will be used. Consider a table having three categorical predictors, A, B, C with categories A_i, B_j, C_k , where $i \leq r, j \leq s$, and $k \leq t$. Let $\pi_{ijk} = P(A_i \cap B_j \cap C_k)$, and let p_{ijk} be the sample estimates of π_{ijk} . The standard summation notation is assumed, ie $p_{jk} = \sum_i p_{ijk}$, $p_{i..} = \sum_{j,k} p_{ijk}$, etc.

Bartlett's definition of no second order interactions can then be written as

$$\frac{p_{111} p_{221}}{p_{121} p_{211}} = \frac{p_{112} p_{222}}{p_{122} p_{212}}. \quad (1.2.1)$$

His method of testing required the solution of a cubic equation. The definition of no interaction in a $2 \times 2 \times 3$ table was also defined, although this required the solution of two simultaneous degree four equations in two variables.

Since then much research has arisen in the study of contingency tables and interactions. Most of the later work, however, can be traced to one of two methods presented in the 1950's. The first method was introduced by Lancaster [24] in 1951, who developed a method of partitioning a χ^2 statistic in order to develop a concept of interaction in the general $r \times s \times t$ table. In the notation above, the hypothesis of no three-way interaction can be written as

$$H_0: \frac{p_{ijk}}{p_{i..} p_{j..} p_{k..}} = \frac{p_{jk}}{p_{j..} p_{k..}} + \frac{p_{ik}}{p_{i..} p_{k..}} + \frac{p_{ij}}{p_{i..} p_{j..}} - 2 \quad (1.2.2)$$

Letting

$$\tilde{p}_{ijk} = p_{i..} p_{j..} p_{k..} \left(\frac{p_{jk}}{p_{j..} p_{k..}} + \frac{p_{ik}}{p_{i..} p_{k..}} + \frac{p_{ij}}{p_{i..} p_{j..}} - 2 \right) \quad (1.2.3)$$

the chi-square statistic for the hypothesis (1.2.2) can be written as

$$\begin{aligned}
 \sum_{ijk} \frac{(p_{ijk} - \tilde{p}_{ijk})^2}{p_{i..} p_{j..} p_{k..}} &= \sum_{ijk} \frac{(p_{ijk} - p_{i..} p_{j..} p_{k..})^2}{p_{i..} p_{j..} p_{k..}} - \sum_{jk} \frac{(p_{jk} - p_{j..} p_{k..})^2}{p_{j..} p_{k..}} \\
 &\quad - \sum_{ik} \frac{(p_{ik} - p_{i..} p_{k..})^2}{p_{i..} p_{k..}} - \sum_{ij} \frac{(p_{ij} - p_{i..} p_{j..})^2}{p_{i..} p_{j..}} \\
 &= \chi^2_{ABC} - \chi^2_{BC} - \chi^2_{AC} - \chi^2_{AB} \tag{1.2.4}
 \end{aligned}$$

Each term on the right hand side of (1.2.4) corresponds to a test of independence of two (or more) predictors. For example, χ^2_{AB} is the χ^2 statistic for the hypothesis that factors A and B are independent. It was these individual terms that Lancaster used to define the interactions between the predictors.

The second approach came in 1956 with the work of Roy and Kastenbaum [28]. Generalizing Bartlett's definition of independence (1.2.1) to the $r \times s \times t$ table they defined interaction in terms of the $(r-1)(s-1)(t-1)$ ratios:

$$\frac{p_{rst} p_{rjt}}{p_{ist} p_{rjt}} / \frac{p_{rsk} p_{rjk}}{p_{isk} p_{rjk}} \left\{ \begin{array}{l} i \leq r-1 \\ j \leq s-1 \\ k \leq t-1 \end{array} \right. \tag{1.2.5}$$

The hypothesis of no interaction was defined as all of the terms being equal to one. They tested the hypothesis of no interaction by the maximum likelihood method, leading to $(r-1)(s-1)(t-1)$ simultaneous equations of degree three. The standard χ^2 formula was then used as a test criterion.

One important method derived from this approach which was initiated by Birch [6] and used by many subsequent authors [5], [7], [22], is the log-linear model. Following Goodman's [22] notation, the probabilities p_{ijk} were written as

$$\log p_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \tag{1.2.6}$$

with the ANOVA-like constraints

$$\left. \begin{aligned} \lambda_i^A = \lambda_j^B = \lambda_k^C &= 0 \\ \lambda_{ij}^{AB} = \lambda_{ik}^{AC} = \lambda_{jk}^{BC} &= \lambda_{ijk}^{ABC} = \lambda_{ikj}^{ABC} = \lambda_{jki}^{ABC} = \lambda_{jik}^{ABC} = \lambda_{kij}^{ABC} = 0 \end{aligned} \right\} \quad \text{for all } i, j, k \quad (1.2.7)$$

The λ 's represented the possible effects of the three variables. The main effects were λ^A , λ^B , and λ^C , the first order interactions were represented by λ^{AB} , λ^{AC} , and λ^{BC} , while λ^{ABC} represented the second order interaction. The model had the advantage of having simple methods for estimating these effects, allowing straight forward calculations for testing of the presence of given interactions and having a obvious extension to arbitrarily large and complex tables.

Assuming the elements in the table represent counts, difficulties arise in these procedures when some of the entries are zero. The behavior of Lancaster's χ^2 statistic is not well understood nor is well-behaved and consequently loses much of its appeal in these cases. More drastic is the behavior of the log-linear model. $\log 0 = -\infty$, creating rather unpleasant consequences in fitting model (1.2.6). Eliminating those categories which contain zero entries loses information and is unreasonable when the number of such entries is of even moderate size.

There has been two main lines of attack to salvage the above mentioned theories. The first and to some extent less successful approaches have dealt with sampling zeroes. One approach considered was to eliminate empty cells or cells having few entries by collapsing adjacent rows of the table. Craig [12] demonstrated a methodology of collapsing two adjacent rows which yielded consistent estimates of the corresponding cell means and a χ^2 square statistic which converged asymptotically to a χ^2 distribution. Bishop [8] approached the problem from the log-linear model and examined conditions under which collapsing the table by adding over a variable would not affect any multi-factor effects. However, for large

tables with many missing cells, the method of collapsing of a table to eliminate all missing entries may not be feasible since it may reduce the table to a size much smaller than desired. Consequently, this method works best only when the number of zero cells is a small fraction of the total number of cells.

Another approach has been to replace the zero entries with some positive number and to carry out the analysis with the altered table. Probably the simplest method along these lines is due to Berkson [4] who suggested replacing 0 with the value 1/2. Variants of this procedure have arisen in the literature, including suggestions of adding either one or one half to all cells in order to eliminate zero entries. A more sophisticated methodology would be to replace the zeroes in the table with the expected values of the cells under some model. Deming and Stephan [13] introduced the iterative proportional fitting procedure for fitting the table with the maximum likelihood estimates of the expected cell frequency. The procedure has the advantage over Berkson's suggestion of being less arbitrary, but relies very much on the parametric model assumed. Both methods may be attacked on a general philosophical ground that analyzing data that has been augmented by addition of values of either an arbitrary nature or by assumption of a specific parametric model should be avoided whenever possible.

There have been several successful strategies proposed for structural zeroes. By and large these have been the classical techniques with minor modifications to account for the zeroes. To simplify the discussion the remainder of the chapter will consider only $R \times C$ tables.

One type of table with structural zeroes which occurred early in the history of contingency tables was separable tables [21]. A table having unordered categories and structural zeroes is said to be separable if it is possible to reorder the categories of the predictors in order to obtain a table having block diagonal form; that is if there exists a partition of the categories of the first predictor $\{A_i\}_{i=0}^n$ and a partition of the categories of the second predictor $\{B_j\}_{j=0}^n$ such that the $\pi_{rs} = 0$ if $r \in A_i$, $s \in B_j$, with $i \neq j$. In such cases the analysis reduces to the analysis of n contingency sub-tables which are asymptotically independent. Standard techniques then can be applied to the individual tables [21].

A second approach, applied to non-separable tables, extending the log-linear model (1.2.6) is the concept of quasi-independence, due to Goodman [21]. Let $S = \{(i, j) \mid \pi_{ij} \neq 0\}$. The probabilities π_{ij} are written

$$\log \pi_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad \text{for } (i, j) \in S \quad (1.2.9)$$

with the constraints

$$\begin{aligned} \sum_i \lambda_i^A &= \sum_j \lambda_j^B = 0 \\ \sum_{(i,j) \in S} \lambda_{ij}^{AB} &= \sum_{(i,j) \in S} \lambda_{ij}^{AB} = 0 \end{aligned} \quad (1.2.10)$$

The model of quasi-independence in this case can be expressed as $\lambda_{ij}^{AB} = 0$ for $(i, j) \in S$. Methods corresponding to the analysis of tables lacking structural zeroes can then be applied (see [14] for example).

Recently several methods have appeared in the literature to estimate the means of large contingency tables which have many empty cells due to sampling variations. These methods are based on the assumption that the categories of the table are ordered so that there is a smooth transition of cell probabilities as one proceeds along any row or column. Fienberg and Holland [16] analyzed sparse tables in a Bayesian framework and considered estimators based on a Dirichlet prior and a squared - error loss function. The smoothness condition was expressed by having the cell means satisfy a functional equation involving a C^2 function of the row and column number.

More recently, Simonoff [29] has adapted the maximum penalized likelihood methodology of density estimation to the estimation of cell means in large sparse contingency tables. Assuming a multinomial likelihood for the cells, the estimates of the cell means are those values which maximize the multinomial likelihood minus some penalty which measures the "roughness" of the estimates. Unfortunately, implementation of the procedure to large tables is difficult and the behavior not well understood in these cases.

One final methodology considered which can be used to relate the behavior of neighboring cells to each other is the method of scoring. This methodology does not appear often in the statistics literature but is not uncommon in other fields. Fisher [17] is usually credited with the first use of scoring. The approach is to assign some values to the categorical predictors in order to maximize the correlation between the predictors. A detailed description can be found in Kendall and Stuart [23]. The approach followed by the psychometricians is different. The philosophy of theirs is closely related to Pearson's ideas of categorical variables in contingency tables. The purpose of optimal scoring, as it is called in the literature, is to assign to the categories of the predictors a value corresponding to some unseen metric. These values, called scores, relate the discrete variables to some continuous measurement. Once assigned, the scored variables are treated as continuous random variables and standard techniques can be applied to them. The term "optimal" comes from the fact that the scores are assigned as to optimize the fit between the scored variables and the model fitted. A laborious and detailed theory of the above may be found in [31].

Chapter 2

The Algorithm

Given a $m_1 \times m_2 \times \cdots \times m_n$ contingency table with Y_{i_1, i_2, \dots, i_n} as the $(i_1, i_2, \dots, i_n)^{\text{th}}$ entry, a relatively flexible model which may be hypothesized is

$$Y_{i_1, i_2, \dots, i_n} = \theta\left(\sum_{j=1}^n S_j(i_j)\right),$$

where S_j are scores and $\theta(\cdot)$ is a “smooth” function. Such a model can be fitted by adapted version of a general procedure called **P.ACE**, which is itself an adaptation of another procedure called **ACE**.

The first section of this chapter briefly describes the **P.ACE** algorithm, as well as its predecessor **ACE**. The second section describes in detail the model and motivates the algorithm used.

2.1. ACE and P.ACE.

Alfréd Rényi [27], in 1959 considered the problem of defining the dependence of two random variables (X, Y) . Consideration of certain natural conditions which he felt such a measure should possess led to the concept of maximal correlation between (X, Y) to be the definition he chose. The maximal correlation between two random variables η and ν is defined as

$$\mathbf{S}(\eta, \nu) = \sup_{f,g} \text{Cor}(f(\eta), g(\nu)),$$

where the sup is taken over all functions for which the left hand side is defined.

In the paper necessary conditions for the existence of two functions f and g such that $\mathbf{S}(\eta, \nu) = \text{Cor}(f(\eta), g(\nu))$ were proved. Such functions can be thought of expressing the natural scale when considering the relationship between (X, Y) since $f(X)$ and $g(Y)$ are in some sense as similar as possible. No attempt was made, however, to establish a method for determining f and g .

In 1982 Breiman and Friedman [10] considered the following generalization of this problem:

Given $X_1, \dots, X_p, Y \in L^2(\Omega, \mathcal{B}, P)$ determine $\varphi_1, \dots, \varphi_p$ and ϑ which minimize $E(\vartheta(Y) - \sum_{i=1}^p \varphi_i(X_i))^2$ with respect to all square integrable functionals subject to the constraint $E(\vartheta^2(Y)) = 1$.

They were able to prove existence and uniqueness of such functions and developed a procedure of determining these functions. Their procedure, christened ACE for Alternating Conditional Expectation, is an iterative procedure yielding a sequence of functions $\{\vartheta^{(i)}, \varphi_1^{(i)}, \dots, \varphi_p^{(i)}\}_{i=1}^\infty$, which converge to the solution under some regularity conditions.

A related problem may be posed: *Given Y, X_1, \dots, X_p as before, determine $\theta, \phi_1, \dots, \phi_p$ which minimize*

$$E\left(Y - \theta\left(\sum_{i=1}^p \phi_i(X_i)\right)\right)^2. \quad (2.1.1)$$

Such a problem may occur when the Y variable is thought of as a response and the vector of X variables are thought of as predictors, and it is desired to predict the value of Y from the X 's. The ACE algorithm yields a prediction rule for the transformed Y , which may not be optimal for predicting the untransformed response. The naïve approach to

this problem is to apply the **ACE** algorithm to determine ϑ , and $\{\varphi\}_{i=1}^p$ which minimize $E(\vartheta(Y) - \sum_{i=1}^p \varphi_i(X_i))^2$ and define $\theta = \vartheta^{-1}$ and $\phi_i = \varphi_i$ for $i = 1, \dots, p$. For noninvertible ϑ this methodology is of little value. Further, given the existence of ϑ^{-1} , this approach in general does not yield the desired result. If θ and $\{\phi_i\}_{i=1}^p$ minimize both

$$E(\theta^{-1}(Y) - \sum_i \phi_i(X_i))^2 \quad \text{and} \quad E\left(Y - \theta\left(\sum_i \phi_i(X_i)\right)\right)^2 \quad (2.1.2)$$

then

$$\theta^{-1}\left(E(Y \mid \sum_{i=1}^p \phi_i(X_i))\right) = E(\theta^{-1}(Y) \mid \sum_{i=1}^p \phi_i(X_i)). \quad (2.1.3)$$

This is in general not true. In particular if θ is either non-linear concave or convex Jensen's inequality prevents (2.1.3) from occurring.

A solution to this prediction problem, known as **P.ACE** for Predictive **ACE**, was proposed by Friedman and Owen [19]. As with **ACE** it is an iterative procedure and is outlined below.

Set $\theta^{(0)} = \varphi_0^{(0)} = \varphi_2^{(0)} = \dots \varphi_p^{(0)} = id$;

Set $i = 1$.

Iterate until convergence of $\theta^{(i)}$, $\phi_0^{(i)}$, $\dots \phi_p^{(i)}$

$\theta^{(i+1)}(Y)$ is a "smooth" of Y on $\sum_{k=1}^p \phi_k^{(i)}(X_k)$

$\phi_j^{(i+1)}$ = those functions which minimize $E\left(Y - \theta\left(\sum_{i=1}^p \phi_i^{(i+1)}(X_i)\right)\right)^2$

$i \mapsto i + 1$

End outer loop iteration

The second step of the inner-loop is implemented by a procedure which alternately updates each $\phi_j^{(i)}$ while fixing the other functions, and continues until each ϕ function

converges. A special case of this method will be used to analyze contingency tables and is discussed in detail in the next section.

2.2. Notation and Description of Algorithm.

The notation for the remainder of the work is now defined. Let $(Y, X_1, \dots, X_p) = (\vec{Y}, \vec{X})$ be random variables on some probability space (Ω, \mathcal{B}, P) . Let X_i be categorical variables, with $X_i: \Omega \rightarrow \{1, 2, \dots, n_i\}$ where $n_i < \infty$ for $i = 1, \dots, p$. Associated with each categorical variable is a scoring with the constraint that the sum of the scores be equal to zero. Let $\mathfrak{p}_j = \{f: N \rightarrow \mathbb{R} \mid \sum_{k=1}^j f(k) = 0\}$. Then every $S \in \mathfrak{p}_{n_i}$ is a scoring for X_i . An additional constraint may be imposed if X_i is an ordered categorical variable. If \prec is the order relation amongst the categories $\{1, \dots, n_i\}$ of X_i then the scores S should satisfy $S(i) < S(j)$ for $i \prec j$.

No assumptions are made concerning the joint distribution of \vec{X} . The model postulated is

$$E(Y|X_1 = x_1, \dots, X_p = x_p) = \sum_{j=1}^m \theta_j \left(\sum_{i=1}^p S_i^j(X_i) \right), \quad (2.2.1)$$

where m is finite and $S_i \in \mathfrak{p}_{n_i}$. $\theta_j(\cdot)$ are elements from a class of functions which are constrained only by a vague concept of "smoothness". Some scaling of the scores is required to insure that the scores and the function are well-defined. The scaling used by the author is $\|\sum S_i(X_i)\|_\infty = 1$.

Two particular instances of this model are now presented. The first example is the model

$$Y = \sum_{j=1}^m \theta_j \left(\sum_{i=1}^p S_i^j(X_i) \right). \quad (2.2.1a)$$

In this case no error is assumed in the formulation of the model and any deviation of the data from the model is due exclusively to sampling errors. Such a model arises when $Y_{i_1, \dots, i_p} = P(X_1 = i_1, \dots, X_p = i_p)$ and the elements of the table are the proportions of counts falling in that position. In this particular case the constraints $Y_{i_1, \dots, i_p} \geq 0$ and $\sum_{i_1, \dots, i_p} Y_{i_1, \dots, i_p} = 1$ are imposed.

On the other hand, if the Y_{i_1, \dots, i_p} are rates corresponding to those elements of the population having characteristics $(X_1 = i_1, \dots, X_p = i_p)$, then a more appropriate model might be

$$Y = \sum_{j=1}^m \theta_j \left(\sum_{i=1}^p S_i^j(X_i) \right) + \epsilon, \quad (2.2.1b)$$

where it is assumed that the ϵ is independent of (Y, \vec{X}) , with $E(\epsilon) = 0$. Although the previous model could be thought of as a special case of this with ϵ representing the sampling error, this is not the intent. In theory, data collected from model (2.2.1b) could have several observations corresponding to the same predictor variables while in model (2.2.1a) at most one observations could ever arise for a particular set of predictor variables.

A particular instance of the random variables (Y, \vec{X}) is a finite set of points in $\Re \times \Re^p$, denoted by $(Y_i, \vec{X}_i)_{i=1}^n = (Y_i, X_{1,i}, \dots, X_{p,i})_{i=1}^n$. It is desired to fit to this instance a model of the form (2.2.1a) or (2.2.1b). For the time being it will be assumed that $m = 1$ so that the outer sum contains exactly one summand. The model is then fitted by choosing scores S_i and a function θ which minimize

$$\sum_{i=1}^n \left(Y_i - \theta \left(\sum_{j=1}^p S_j(X_{j,i}) \right) \right)^2. \quad (2.2.2)$$

A variant of the **P.ACE** algorithm mentioned in the previous section is used to fit the model. The procedure iteratively updates the function $\theta(\cdot)$ and the scores S_j until convergence. For the scores fixed $\theta(\cdot)$ is determined by “smoothing” the data Y_i on the sum of the scores $\sum_{j=1}^p S_j(X_{j,i})$. Readers unfamiliar with smoothing procedures will find a short discussion in the appendix.

For $\theta(\cdot)$ fixed, determining the scores S_j is a large non-linear minimization problem. A direct attack is expensive and consequently a simple “approximate” solution is chosen. It is sufficient to consider optimizing $S_1(X_1)$. Let $I_i = \{j : X_{1,j} = i\}$. Then

$$\sum_i \{(Y_i - \theta(\sum_j S_j(X_{1,j})))\}^2 = \sum_i \sum_{I_i} \{(Y_i - \theta(\sum_j S_j(X_{1,j})))\}^2. \quad (2.2.3)$$

Consequently, to minimize the left hand side it is sufficient to minimize each summand on the right hand side. From symmetry it is evident that only one term, which for definiteness sake is taken to be I_1 , is needed to be analyzed in detail. There are two approaches which will yield the desired result. The first method is a simple application of calculus and is left as an exercise for the reader. The second method is more convoluted and obscure. It is the second method which will be illustrated, not because of any perversity of the author but rather because it illustrates the original P.ACE procedure, and yields a result that is easily interpreted.

Let S_j be some set of scores, $w_i = \sum_j S_j(X_{1,j})$, S_1^1 be the value of the optimal score for the first predictor variable at the point 1, and $\delta = S_1^1 - S_1(1)$. The problem thus reduces to finding the value of δ which minimizes

$$\sum_{I_1} (Y_i - \theta(w_i + \delta))^2 \quad (2.2.4)$$

Assume $\theta(\cdot) \in C^2$ so that $\theta(w_i + \delta) = \theta(w_i) + \delta \theta'(w_i) + O(\delta^2)$. For each i let δ_i^* to be that value which minimizes $(Y_i - \theta(w_i + \delta_i^*))^2$. Then $\sum_{I_1} (Y_i - \theta(w_i + \delta_i^*))^2$ is minimized over all sets of numbers $\{\delta_i^*\}_i$, and all that remains is to select a single value for δ . To simplify notation let $\epsilon_i = Y_i - \theta(w_i + \delta_i^*)$, and $m_i = \theta'(w_i + \delta_i^*)$.

$$\begin{aligned} \sum_{I_1} (Y_i - \theta(w_i + \delta))^2 &= \sum_{I_1} (Y_i - \theta(w_i + \delta_i^*) + \theta(w_i + \delta_i^*) - \theta(w_i + \delta))^2 \\ &= \sum_{I_1} (\epsilon_i + \theta(w_i + \delta_i^*) - [\theta(w_i + \delta_i^*) + (\delta - \delta_i^*) m_i + O(\delta - \delta_i^*)])^2 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{I_1} (\epsilon_i + (\delta - \delta_i^*) m_i + O(\delta - \delta_i^*))^2 \\
 &= \sum_{I_1} \epsilon_i^2 + 2 \sum_{I_1} \epsilon_i m_i (\delta - \delta_i^*) + \sum_{I_1} (\delta - \delta_i^*)^2 m_i^2 + O((\delta - \delta_i^*)^2) \\
 &\stackrel{?}{=} \sum_{I_1} \epsilon_i^2 + 2 \sum_{I_1} \epsilon_i m_i (\delta - \delta_i^*) + \sum_{I_1} (\delta - \delta_i^*)^2 m_i^2 \quad (a) \\
 &= \sum_{I_1} \epsilon_i^2 + \sum_{I_1} (\delta - \delta_i^*)^2 m_i^2 \quad (b)
 \end{aligned}$$

Equation (a) results from a flagrant abuse of $O(\cdot)$ notation, whereas equation (b) is a valid consequence of the fact that $\epsilon_i m_i = 0$ since δ_i^* minimizes $((Y_i - \theta(w_i + \delta_i^*))^2$. It is clear that to minimize this it is necessary and sufficient that $0 = \sum_{I_1} (\delta - \delta_i^*) m_i^2$, or equivalently

$$\delta = \left(\sum_{I_1} \delta_i^* m_i^2 \right) / \sum_{I_1} m_i^2 \quad (2.2.5)$$

It is noted that the optimal “global” increment is a weighted average of the optimal “local” increments. For m_i^2 small, a small deviation from δ_i^* and consequently a small change in the function $\theta(\cdot)$ results in a small increase of the squared error. Conversely for m_i^2 large, a small deviation from δ_i^* results in a large increase of the squared error. Consequently, a rational procedure would down-weight those values for which m_i^2 is small and give greater weight to those values for which m_i^2 is large. The updating algorithm (2.2.5) does precisely this.

Apart from ease of interpretation this solution is less than optimal for updating the scores. The determination of δ_i^* is time consuming, appearing directly in the numerator and indirectly in the denominator of (2.2.5). To simplify even further, the ubiquitous Taylor’s theorem is invoked twice. Without loss of generality it may be assumed that $m_i \neq 0$. Hence $0 = Y_i - \theta(w_i + \delta_i^*) = Y_i - \{\theta(w_i) + \delta_i^* \theta'(w_i) + O([\delta_i^*]^2)\}$, which with $m_i = \theta'(w_i + \delta_i^*) = \theta'(w_i) + O(\delta_i^*)$ yields $Y_i - \theta(w_i) = m_i \delta_i^* + O([\delta_i^*]^2)$. Substituting these two expressions into (2.2.5) and after some sleight of hand involving the further misuse of $O(\cdot)$ one arrives at

$$S_1^1 = S_1(1) + \sum_{I_1} [Y_i - \theta \left(\sum_{j=1}^p S_j(X_{i,j}) \right)] \theta' \left(\sum_{j=1}^p S_j(X_{i,j}) \right) / \sum_{I_1} [\theta' \left(\sum_{j=1}^p S_j(X_{i,j}) \right)]^2 \quad (2.2.6)$$

Despite its appearance, this updating algorithm is easily implemented. The term $Y_i - \theta \left(\sum_{j=1}^p S_j(X_{i,j}) \right)$ is merely the residual of the model from the observation and can be computed when calculating $\theta'(\cdot)$. As explained in the appendix $\theta' \left(\sum_{j=1}^p S_j(X_{i,j}) \right)$ likewise can be estimated at this time. The difference between the solution (2.2.6) and the solution of (2.2.5) is of order $O(\delta - \delta_i^*)$ which is small when the algorithm is close to the true minimizing functions.

No order constraints for the scores corresponding to ordered categorical variables have yet been imposed. If alternately smoothing and applying the updating algorithm (2.2.6) yields as the result scores with the correct order, then the constrained problem coincides with the unconstrained problem and no difficulty arises. If, on the other hand, this is not the case the offending scores must be coerced into submission. This is accomplished by use of the pool adjacent violators algorithm. The method is discussed in detail in [2]. Since the order of the scores the procedure finds yields information about the data this approach of forcing a particular ordering of the scores will not be considered here.

The algorithm is then continued with the new scores and categories until convergence. The procedure of pooling the scores for ordered variables if necessary and continuing the algorithm continues until the procedure converges with the scores of ordered categories obeying the order constraints.

Thus a method of determining scores and a function to minimize (2.2.2) is determined and all that remains is to extend this method for determining functions and scores to minimize

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^N \theta_j \left(\sum_{k=1}^p S_k^{(j)}(X_{k,j}) \right) \right)^2, \quad (2.2.7)$$

where N is some small number.

To this end a greedy algorithm is employed in conjunction with the procedure outline above. Let $\{\theta_1, S_1^{(1)}, \dots, S_p^{(1)}\}$ be the function and scores obtained by minimizing (2.2.2). Define a new set of response variables $\hat{Y}_i = Y_i - \sum_{j=1}^p \theta_1(\sum_{k=1}^n S_k^{(1)}(X_{k,j}))$, for $i = 1, \dots, n$. The second set of functions and scores are then determined according to the methodology described above using the residuals from the first fit as the new response variable. This procedure of replacing the response variable with the residuals from the previous fit and finding the minimizing function and scores for (2.2.2) yields the sequence of functions and scores, and is repeated until it is deemed that no acceptable improvement in the model is obtained. In practice, however, it has been found in the examples tried so far that one summand is sufficient to capture most of the structure in the data.

Further modifications can be made to the algorithm. The most obvious corresponds to the method of obtaining the smooth functions $\theta(\cdot)$. Various smoothing algorithms are present, although experience has shown this to have minor impact. Robust smoothers can be utilized, as well as monotone smoothers, or requiring the smooths to lie in some parametric family. Except for a brief discussion in the following chapter these issues will not be addressed in any detail, and the reader is invited to implement any or all of the above, according to his/her indiscretion. The actual smoother implement in the examples included here is a variable span smoother having three separate band-widths, each being fitted to the data by a local least-squares algorithm [18].

To summarize, the algorithm used to fit (2.2.1a) and (2.2.1b) is as follows:

Set $i = 1$

While sum of residuals squared shows large enough decrease

$$\theta_i^{(0)} = S_{1,i}^{(0)} = \dots = S_{p,i}^{(0)} = id$$

Set $k = 1$

Iterate until (2.2.2) fails to decrease

$$\theta_i^{(k+1)}(Y) = \text{smooth of } Y \text{ on } \sum S_{j,i}^{(k)}(X_j)$$

$S_{j,i}^{(k+1)}$ updated according to (2.2.5)

End innermost loop

Set $\theta_i = \theta_j^{(k)}$, $s_{j,i} = s_{j,i}^{(k)}$

Set $Y = Y - \sum s_{j,i}(X_j)$

$i \mapsto i + 1$

End outer loop

Chapter 3

Some comparisons

Whenever a new method is suggested which either extends or generalizes existing procedures an important question is its performance in cases when the previous procedures yield satisfactory results. If the new method fails in these cases or yields results differing significantly from the previously found results, some question of value is raised. It is clearly not possible to examine all cases in which the classical procedures have analyzed tables successfully, and yet reasonable performance on just a few cases is sufficient to establish some basis of trust in it, especially if in these cases the results of the old and the new are comparable. In what follows, several "real" data sets are extracted from various sources for which the classic model seems appropriate and the results are compared to the results of the **P.ACE** algorithm.

As described in the previous section, different results for **P.ACE** may be obtained by use of different smoothers. In particular, a type of parametric **P.ACE** may be obtained if the smooths are constrained to lie in some parametric family. As stated earlier, this parametric **P.ACE** procedure is not advocated. However, the relationship between **P.ACE** and some of the classic procedures is most easily seen by use of this parametric **P.ACE** procedure as an intermediary. In four examples a classical model is shown to be related to the **P.ACE** procedure in which the curve is forced to lie in some parametric family. If the range of the non-parametric smoother of the **P.ACE** procedure encompasses this parametric family then it can be deduced that the **P.ACE** procedure generalizes some variant of the classical procedure.

3.1. Exponential models.

The first family is a two parameter family of exponentials, $F_E = \{ae^{bx} | a, b \in \mathbb{R}\}$. The parameter b is included because of the scale constraint on the scores. To simplify notation assume the table under consideration is an $r \times s \times t$ table. Letting

$$\left\{ \begin{array}{l} \log p_{1i} = b S_1(i) \\ \log p_{2j} = b S_2(j) \\ \log p_{3k} = b S_3(k) \end{array} \right\} \quad (3.1.1)$$

the model corresponding to this family is

$$Y_{ijk} = a \exp\{b(S_1(i) + S_2(j) + S_3(k))\} = a p_{1i} p_{2j} p_{3k} \quad (3.1.2)$$

where p_{1i} , p_{2j} , and p_{3k} satisfy $\sum_i \log p_{1i} = \sum_j \log p_{2j} = \sum_k \log p_{3k} = 0 \quad \forall i, j, \text{ and } k$ and are chosen to minimize

$$\sum (Y_{ijk} - a p_{1i} p_{2j} p_{3k})^2 \quad (3.1.3)$$

This corresponds to the model of independence and is closely related to Birch's method of analyzing contingency tables having no higher order interactions [6]. This method of estimation consists of taking logarithms of the response variable and applying standard ANOVA constraints to the transformed data. In the case of independence this reduces to the model

$$\log Y_{ijk} = \gamma_{1i} + \gamma_{2j} + \gamma_{3k} + c \quad (3.1.4)$$

where $\sum_i \gamma_{1i} = \sum_j \gamma_{2j} = \sum_k \gamma_{3k} = 0 \quad \forall i, j, k$. When the estimates are given by

$$\gamma_{jk} = \sum_{i_j=k} \log Y_{i_1 i_2 i_3} / \sum_{i_j=k} 1 \quad (3.1.5)$$

they can be shown to minimize

$$\sum (\log Y_{ijk} - \gamma_{1i} - \gamma_{2j} - \gamma_{3k} - c)^2. \quad (3.1.6)$$

Another method to fit this model is to use the maximum likelihood estimates. In the case of independence the estimates are given by

$$\gamma_{jk} = \log \left(\sum_{i_j=k} Y_{i_1 i_2 i_3} / \sum_{i_j=k} 1 \right) \quad (3.1.7)$$

It is seen that using the first method of estimation, models (3.1.2) and (3.1.4) differ only in that model (3.1.4) minimizes on a log transformed scale while the proposed model (3.1.2) minimizes on the original scale. Consequently, in the case where model (3.1.4) is appropriate one might expect that they yield similar results. The estimates given by (3.1.5) can be considered as the average of the log of the responses while the estimates given by (3.1.7) are the log of the average of the responses. Although Jensen's inequality forces the estimates to differ, if the log is semi-linear in the region of interest the estimates should be similar, and consequently the the results of models (3.1.2) and (3.1.6) should be similar.

As an example a data set was analyzed using the P.ACE algorithm , the Birch method, and the maximum likelihood method. The data consisted of 60 observations taken from the U.N. Demographic Yearbook 1980 [30]. The response variable was German infant mortality rate and the predictors consisted of four categorical variables described on page 33. Also listed on that page are the scores derived from the Birch analysis (*labeled Anova*), from the maximum likelihood method, (*labeled M.L.*), and from the P.ACE procedure (*labeled PACE*).

The following two pages contain plots of the corresponding curves. On page 34 the sum of scores corresponding to the Anova analysis is plotted in plusses against the response

variable and the corresponding exponential curve is drawn through using dotted lines. Also on the graph are the sum of scores corresponding to the PACE procedure plotted in circles with the estimated curve plotted in solid lines. On page 35 the curve corresponding to the P.ACE analysis is reproduced. The sum of scores of the maximum likelihood procedure is plotted as plusses and the dotted line represents the corresponding exponential curve. As can be seen from the following pages the scores are all very similar to each other, and the curves corresponding to them are nearly identical.

Since all three curves are monotone increasing, large scores correspond to increased mortality rates while small scores correspond to decreased mortality rates. In all three procedures the ordering of the scores within each category is the same, indicating agreement of the procedures in the relative ranking of attributes related to increase mortality. The largest and smallest scores occur in the category of age at death, indicating that this is the most important factor in determining early and late mortality. The period of elapsed time involved in the different age groups vary, making direct comparisons difficult. It should be noted that the two periods of smallest time elapsed (less than one day and between one and six days) have the largest scores, indicating that these periods of time are at greater risk than any other period involved, while the category consisting with the largest amount of elapsed time (five to eleven months) has the smallest scores, indicating the smallest risk group. Both country of birth and sex of child follow as important factors, with scores of approximately equal magnitudes. The scores corresponding to the year of birth are very small, indicating that year of birth is the least informative category.

3.2. Linear models.

The second family consisted of the two parameter family of linear functions, $F_L = \{ax + b \mid a, b \in \mathbb{R}\}$. The parameters a and b are required because of the scaling of the scores. The model corresponding to this family is

$$Y_{i_1, \dots, i_p} = a \sum S_j(i_j) + b \quad (3.2.1)$$

Such a model can be thought of arising from taking logarithms of (3.1.2) with the response in (3.2.1) corresponding to the logarithm of the response in (3.1.2). The linear model, however, arises naturally in several untransformed models. This particular restriction of the smoother corresponds exactly to a procedure described by Young [31]. The method suggested for fitting such models was to alternate between optimizing the scores and the linear fit, a procedure corresponding exactly in spirit, if not in detail, to the **P.ACE** procedure. Consequently, **P.ACE** can be thought of as a direct generalization of this method.

As an example a data set was analyzed using the **P.ACE** algorithm in which the smooth functions were constrained to be linear and the unmodified **P.ACE** procedure. The data consisted of 140 observations taken from the U. N. Demographic Yearbook [30]. The response variable was expected number of years prior to death and the predictors consisted of the nationality, age, and sex of the subject as well as the year of the estimate. These can be found on page 36, as well as the scores derived from both models.

On the following page the sum of the scores corresponding to the linear model is plotted in plusses against the response variable and the corresponding straight line is drawn using dotted lines. Also on the same graph are the scores corresponding to the unmodified **P.ACE** procedure drawn in circles and the corresponding curve drawn using a solid line. As can be seen from the graph and the table, the two sets of scores are very similar and not surprisingly the curves are also quite similar.

Large scores correspond to a decrease in life expectancy, while small scores correspond to an increase. With one exception (*Cuba versus Italy*) the ordering within each category is the same, indicating that both procedures agree in the relative effect and importance of the different categories. As was to be expected, age was the most important factor. The next overall important variable was sex, indicating that females tended to have a definite advantage over males in terms of life expectancy. Much less important was the year of the estimate, and almost irrelevant was the country. Since it was the country variable which

displayed the difference in ordering, not too much importance should be placed on this aberration. Overall, the fit seems quite close.

3.3. Box-Cox models.

A family of functions of transformations was introduced by Box and Cox in 1974 [9]. The family, now known as Box-Cox transformations, is a one parameter family of curves defined by $F_{CB} = \{f_\lambda(x) = x^\lambda \mid \lambda \in \mathbb{R}\}$, with the condition that $f_0(x) = \lim_{\lambda \rightarrow 0} f_\lambda(x) = \ln(x)$. Given a sequence of predictor and response variables $\{(\vec{x}_i, y_i)\}_{i=1}^N$ it was suggested that a linear fit be found to the transformed sequence $\{(\vec{x}_i, f_\lambda(y_i))\}_{i=1}^N$, where λ is chosen to maximize the fit. Thus the model fitted is

$$f_\lambda(Y) = a_0 + \sum a_j x_j \quad f_\lambda \in F_{CB} \quad (3.3.1)$$

This suggests that an appropriate class of parametric curves to consider in the P.ACE methodology may be the three parameter family of scale and location shifts of the inverses of the Cox-Box transformations, namely $F_{CB}^{-1} = \{f_{\lambda^*}^{-1}(a x + b) \mid a, b \in \mathbb{R}, f_{\lambda^*}(\cdot) \in F_{CB}\}$. It should be noted that this family contains a large class of functions, including both the linear and exponential curves considered earlier. The model corresponding to this family of curves can be written as

$$Y_{i_1, \dots, i_p} = f_{\lambda^*}^{-1}(a + b \sum_{j=1}^p S_j(x_{i,j})) \quad f_{\lambda^*} \in F_{CB} \quad (3.3.2)$$

Clearly the two models are closely related. Superficially it appears possible to transform one into the other by merely inverting the function $f_\lambda(\cdot)$. By the argument in chapter two, however, it is clear that one would not expect the function $f_\lambda(\cdot)$ of (3.3.1) to be the

inverse of the function, $f_{\lambda}^{-1}(\cdot)$, of equation (3.3.2). Except in cases where the function is linear the two functions correspond to different values of λ , although in many cases it would seem reasonable that the two values of λ be close.

When there are only a few distinct values of each predictor it is possible to consider the predictors to be categorical. Writing

$$a_0 + \sum a_j x_j = (a_0 + \sum a_j \bar{x}_j) + b' \sum \left\{ \frac{a_j(x_j - \bar{x}_j)}{b'} \right\} \quad (3.3.3)$$

where b' is included to satisfy whatever prerequisite scaling is being invoked for scores, it can be seen that the centered and scaled predictors $a_j(x_j - \bar{x}_j)/b'$ correspond to scores in the categorical model (3.3.2). There is a difference, however, between the scores of (3.3.2) and the "scores" of (3.3.3). Whereas the scores of the P.ACE procedure are chosen as to maximize the fit to the model and are subjected only to scale and location constraints, the scores of (3.3.3) are constrained to be scale and location shifts of the values of the original predictors, reducing the amount of flexibility in the model. Letting $a' = a_0 + \sum a_j \bar{x}_j$ and $S_j'(x_i) = a_j(x_{j,i} - \bar{x}_j)/b'$, the Box Cox model can be written as

$$f_{\lambda}(Y) = a' + b' \sum S_j'(X_j) \quad (3.3.4)$$

making the relationship to the parametric version of P.ACE transparent.

With this identification it is possible to compare the two procedures on a data set. The data examined comes from the Box Cox paper [9]. The data resulted from attaching to yarn of three different lengths three different sets of weights and subjecting the yarn to swings of three different amplitudes. The response was the number of swings prior to the breaking of the yarn. Since the number of distinct values of the predictor variables is small, it is possible to consider them as categorical and use the P.ACE procedure on them. From the discussion above, however, it can be seen that the resulting scores from the two procedures can not be compared, making it difficult to evaluate the two procedures. To eliminate this problem the Cox-Box model was replaced by the model (3.3.2). This

replacement allowed the same degree of freedom in determining the scores and had both models minimize the squared error between the fit and the data in the same metric. Given that the Box-Cox model and the model (3.3.2) are similar, this approach does not appear to be entirely unreasonable.

The scores of the two procedures, as well as the predictors and the true and estimated values of the response of both models can be found on page 38. On page 39 the sum of scores corresponding to the P.ACE procedure is plotted in circles against the response variable and the fitted line is drawn using a solid line. The sum of scores corresponding to the pseudo-Box-Cox model is plotted in plusses against the response on the same graph and the corresponding curve is plotted in a dotted line.

Note that the curve and the estimates agree rather closely. The curve chosen to maximize the fit was $f(x) = x^{-14.9}$. The inverse of this is the function $f(y) = y^{-0.067}$, which is in close agreement with the transformation Box and Cox found in their paper, namely $f(y) = y^{-0.06}$. Two of the three predictors, length and load, show very good agreement in the scores, and in all cases the ordering is the same.

3.4. Logistic models.

The four parameter family of curves $F_{Lo} = \{a + b e^{cx+d}/(1 + e^{cx+d}) | a, b, c, d \in \mathbb{R}\}$ encompass a large number of curves not contained in the families previously considered. When the response variable represents the probability, P , of the occurrence of some event given a specific set of covariates x_1, \dots, x_k , a model often used is the logistic model,

$$P(x_1, \dots, x_k) = \frac{e^{b+\sum a_i x_i}}{1 + e^{b+\sum a_i x_i}} \quad (3.4.1)$$

corresponding to the sub-family of F_{Lo} in which the additive term is zero and the multiplicative term is one. The reason for the popularity of the model (3.4.1) is due to an

equivalent formulation,

$$\ln \frac{P(x_1, \dots, x_k)}{1 - P(x_1, \dots, x_k)} = \sum a_i x_i + b \quad (3.4.2)$$

which allows a linear analysis of the transformed response. As noted previously, such an analysis will not be optimal when predicting responses from the covariates and in such cases the model fitted should be (3.4.1).

Often the covariates are measurements of some continuous variable. However, as in the case of the Box-Cox models, if there exists only a few distinct values of the covariates they can be considered categorical. In such cases the linear term in (3.4.1) and (3.4.2) is replaced by a location and scale shift of sums of scores.

To demonstrate an instance in which both the logistic model and the P.ACE procedure seem to be in some agreement the win-loss record of the American baseball league in 1948 was examined [1]. The predictor variables consist of each of the eight teams in the American league and the response for the i^{th} j^{th} entry of the table is the number of games team i won against team j that season. This data is an example of a table having structural zeros along the main diagonal, since presumably a team does not play against itself. Initially there appears 56 observations, a pair of wins and losses corresponding to each pair of teams. However, with four exceptions each pair of teams played 22 games, effectively reducing the number of observations to 28. Given such a symmetry it would be reasonable to expect any model based on the data to reflect this symmetry as well.

A straight forward application of the P.ACE methodology would construct a model of the form

$$\# \text{ of wins of team A against team B} = \theta(S_{win}(A) - S_{loss}(B)). \quad (3.4.3)$$

In this model each team would have two scores, one score for the games it won and one score for the games it lost. A reasonable alternative would be to fit the model

$$\# \text{ of wins of team A against team B} = \theta(S(A) - S(B)) \quad (3.4.4)$$

which assign a single score to each team. This model would not only reflect the symmetry of the problem but also reduces the number of estimated parameters by 8. It is this later model that is fitted. This is accomplished by fitting model (3.4.3) with the additional constraint that $S_{win}(A) = -S_{loss}(A)$.

The logistic model fitted was of the form

$$\# \text{ of wins of team A against team B} = \alpha + \beta \times L(S(A) - S(B)) \quad (3.4.5)$$

where L is some scale and location shift of the standard logistic curve and the numbers α and β are chosen to maximize the fit. In both models the scores can be thought of representing some measure of the strength of the teams while the functions represent the means to translate the relative strength two teams into the expected number of wins for either team.

The observed values and the estimates from both models may be found on page 40, as well as the scores obtained by both methods. Page 41 contains the graph of the sum of scores of the P.ACE procedure versus the observed values plotted in circles with the curve estimated by the procedure drawn in a solid line. On the same graph the sum of scores of the logistic fit versus the observed values are plotted in plusses with the corresponding logistic function plotted in a dotted line. The two curves are remarkably similar. The scores are comparable as are both fits to the data. The ordering of both scoring systems is the same and coincides with the final standings of the league at the end of the season, reinforcing the interpretation that the scores represent some measure of relative strength of the teams.

3.5. Comments.

Four different parametric models were considered in this chapter: the exponential model, the linear model, models arising from Box-Cox transformations and the logistic model. In each case the **P.ACE** model was shown to generalize to some variant of the parametric model. For each of the parametric models a data set was selected for which that particular model fitted well. However, no single parametric model was appropriate for all four data sets. On the other hand, the **P.ACE** model fitted all four sets well, yielding results similar to those obtained by the corresponding parametric models. Thus the **P.ACE** procedure has the advantage of freeing the investigator from having to postulate possibly wrong parametric model assumptions without the concern of possibly obtaining very different results if a particular parametric model is appropriate. Because of this flexibility the **P.ACE** model may be used as a model selection procedure directing the researcher towards a specific parametric model.

On the other hand, it is reasonable to believe that data sets exist for which none of the standard parametric models are appropriate while the general **P.ACE** model fits well, although the author has yet encountered such a set. Thus the **P.ACE** procedure may be viewed as a method of fitting a non-parametric model of a contingency table without further reference to any parametric model. In addition, an advantage of the procedure is that it appears to be relatively insensitive to missing values in the table. This aspect is discussed in further detail in the next chapter.

3.6. Graphs and Tables.

This section contains graphs and tables of the data considered in this section. As a matter of consistency the following conventions have been followed. Whenever the **P.ACE** procedure is compared with an alternate procedure

- 1) The sum of scores obtained by the **P.ACE** procedure versus the observations are

plotted as circles.

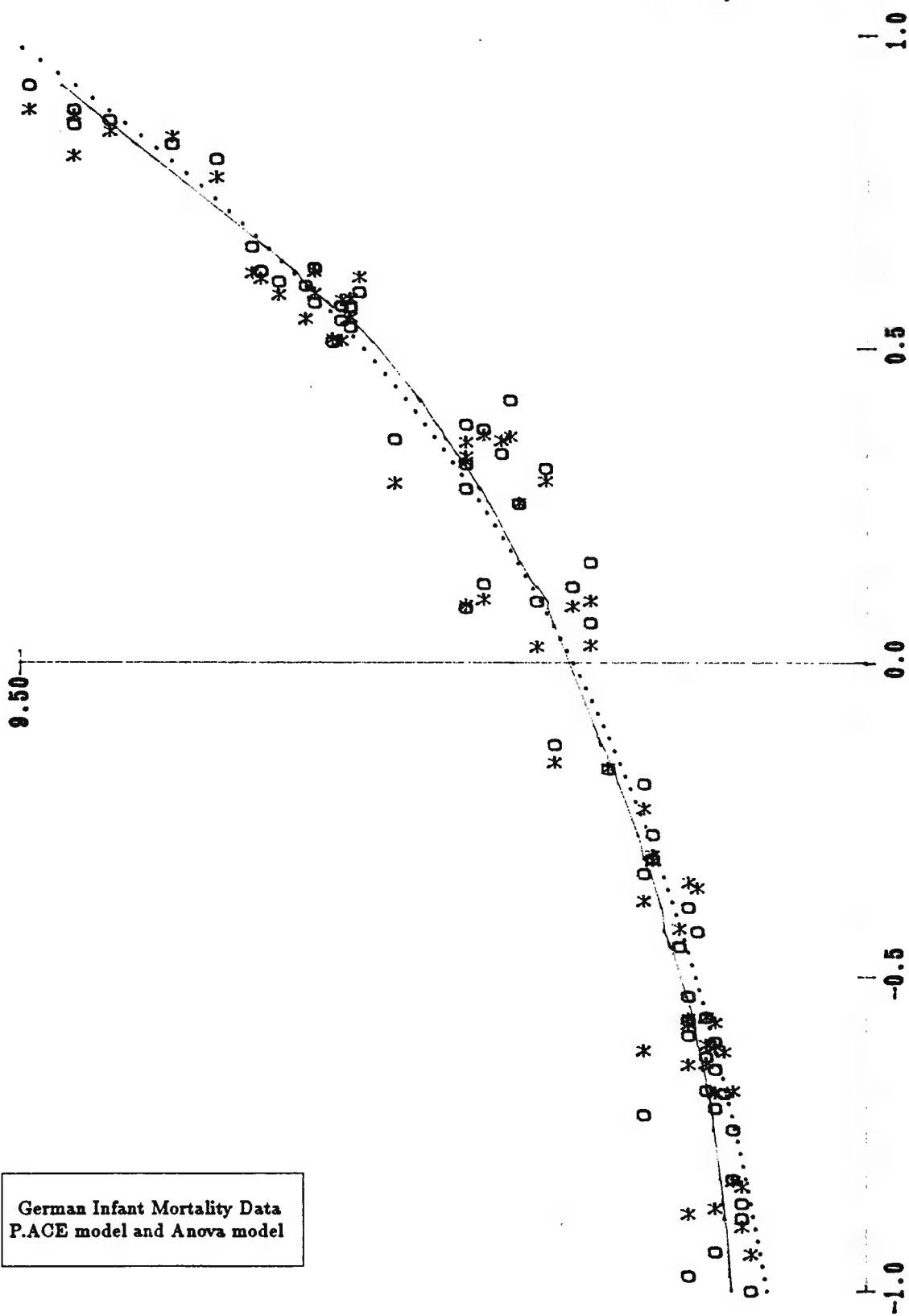
- 2) The curve obtained by the P.ACE procedure is plotted as a solid line
- 3) The sum of scores obtained by the alternate procedure versus the observations are plotted as plusses.
- 4) The parametric curve obtained by the alternate procedure is plotted as a dotted line

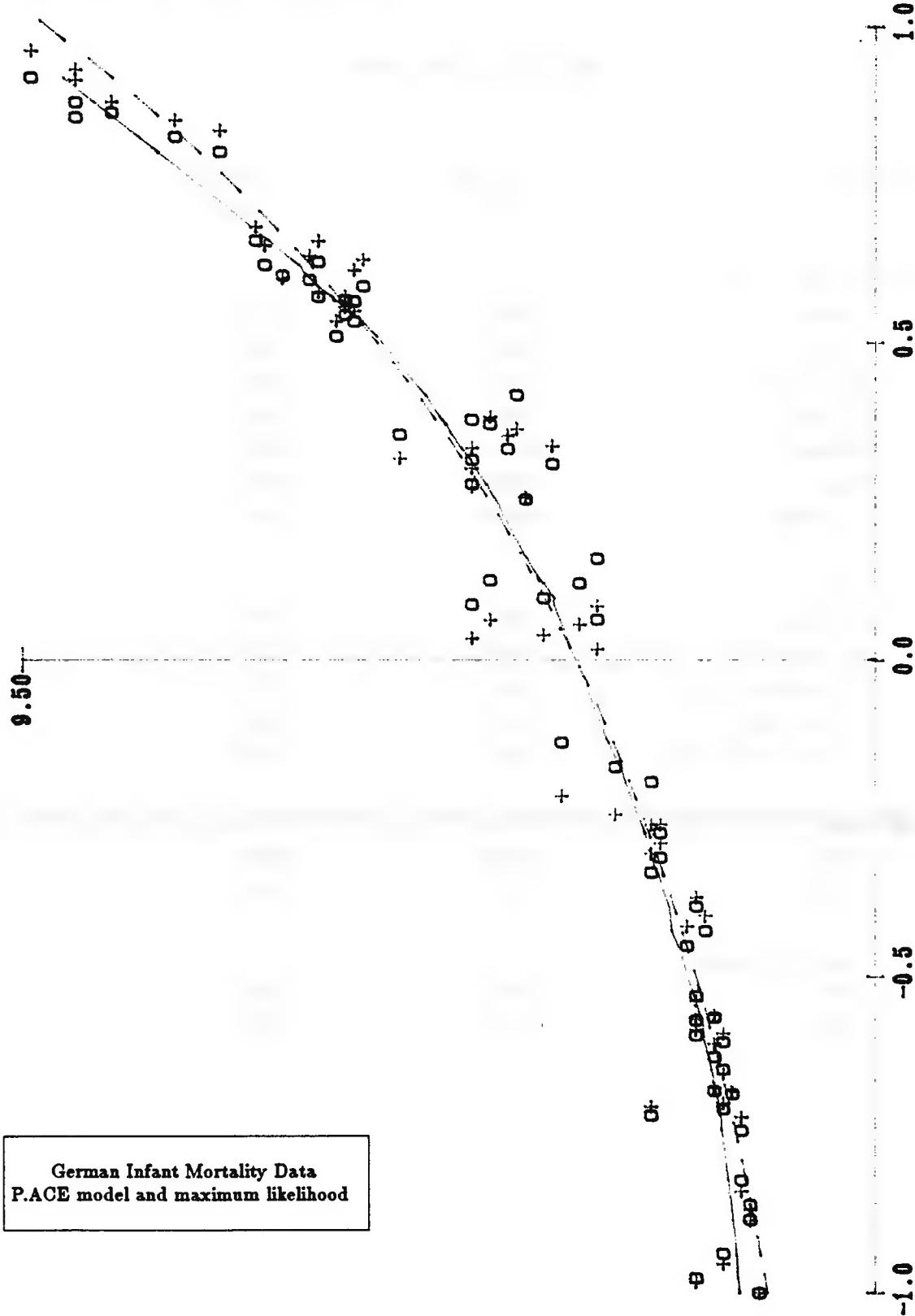
German Infant Mortality Data**Predictor variables**

1) country of birth	0	East Germany
	1	West Germany
2) year of death	0	1971
	1	1972
	2	1973
3) sex of infant	0	male
	1	female
4) age at death	0	less than one day
	1	less than six days
	2	less than twenty seven days
	3	less than five months
	4	less than eleven months

Scores

predictor	M.L.	Anova	P.ACE
1	- 0.150	- 0.129	- 0.161
	0.150	0.129	0.161
2	0.025	0.027	0.036
	- 0.004	0.018	0.006
	- 0.021	- 0.046	- 0.042
3	- 0.139	- 0.130	- 0.152
	0.139	0.130	0.152
4	0.567	0.560	0.561
	0.648	0.595	0.651
	- 0.575	- 0.592	- 0.559
	0.050	0.073	- 0.069
	- 0.690	- 0.636	- 0.584

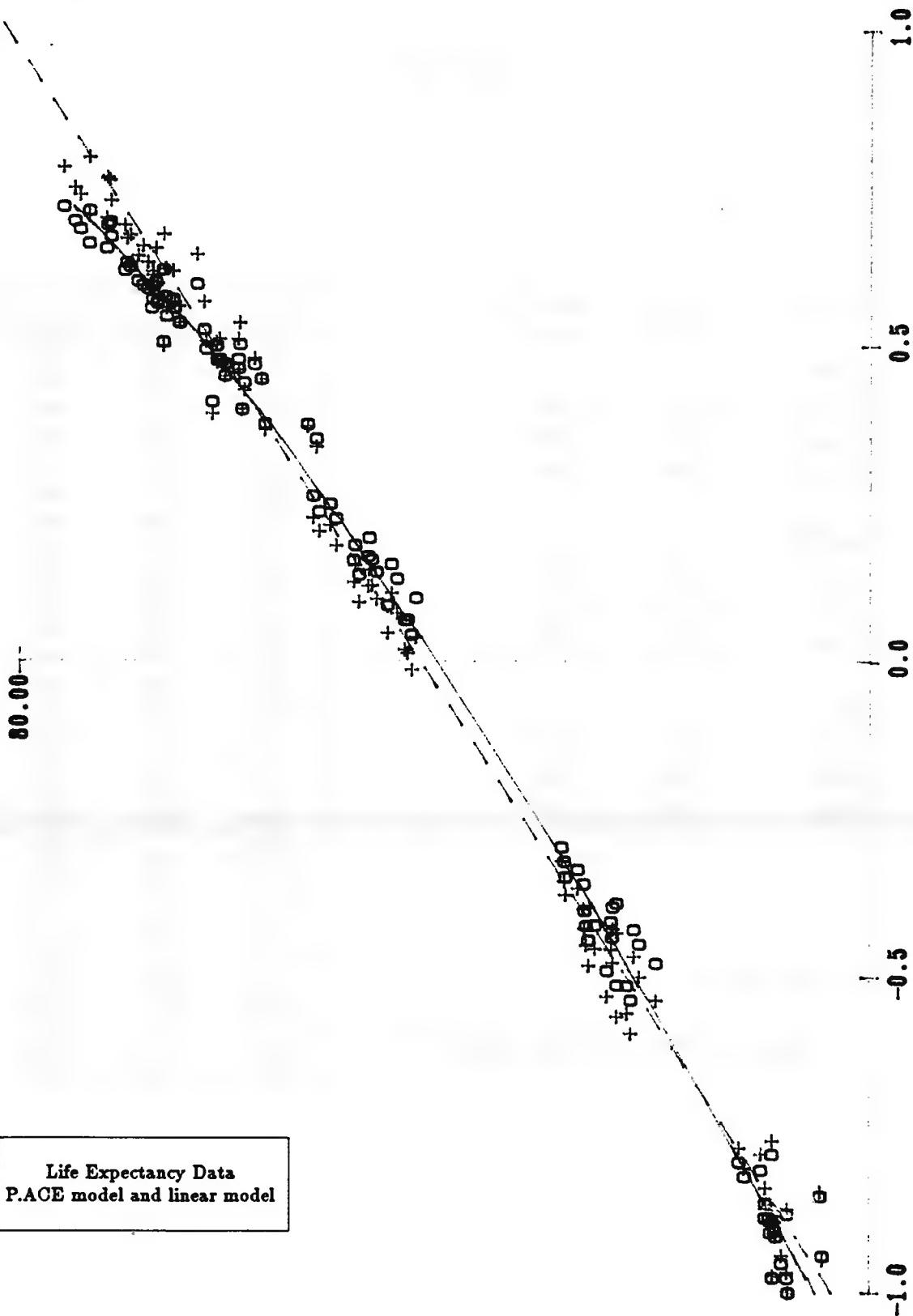




German Infant Mortality Data
P.ACE model and maximum likelihood

Life Expectancy Data

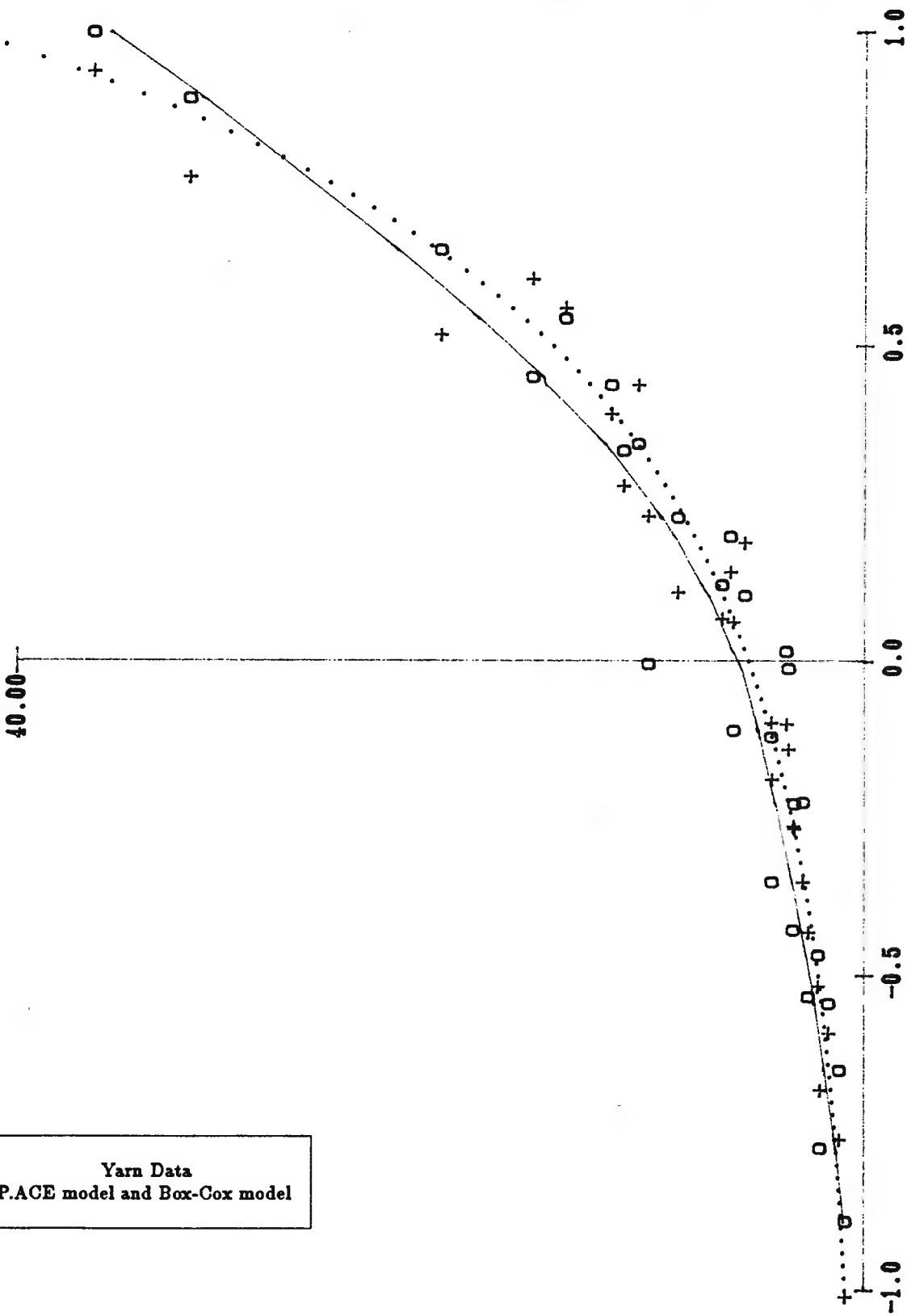
predictor	Linear scores	P.ACE scores
1) <u>country</u>		
Chile	− 0.081	− 0.075
Cuba	0.069	0.049
Singapore	− 0.028	− 0.019
Kuwait	0.032	0.028
Mexico	− 0.054	− 0.051
Italy	0.053	0.053
Scotland	0.009	0.015
2) <u>age</u>		
at birth	0.661	0.627
at ten years	0.490	0.489
at twenty-five years	0.137	0.162
at fifty years	− 0.439	− 0.420
at seventy-five years	− 0.849	− 0.862
3) <u>sex</u>		
male	− 0.542	− 0.502
female	0.542	0.502
4) <u>year of estimate</u>		
1970	− 0.162	− 0.129
1975	0.162	0.129



Yarn Data

Predictor	P.ACE scores	Box-Cox scores	actual values	pace estimates	Box-Cox estimates
Length			0.90	1.08	0.91
250 mm	-0.434	-0.454	1.18	1.73	1.37
300 mm	-0.011	0.039	1.70	2.10	1.85
350 mm	0.444	0.415	2.10	1.39	1.60
			2.20	2.81	2.14
			2.66	2.15	2.46
			2.92	3.40	2.89
Amplitude			3.32	2.38	3.32
8 mm	0.405	0.330	3.38	3.09	3.35
9 mm	-0.144	-0.001	3.60	6.16	4.28
10 mm	-0.261	-0.329	3.70	6.31	4.52
			4.38	3.59	3.91
			4.42	4.90	4.55
Load			5.66	6.20	8.04
40 gm	0.151	0.193	6.20	5.17	6.17
50 gm	0.045	0.024	6.34	7.68	7.34
60 gm	-0.195	-0.217	6.74	6.12	6.24
			8.84	8.22	6.77
			10.22	5.95	8.58
			10.70	11.24	12.97
			11.40	10.77	9.42
			11.98	13.38	11.82
Box - Cox model:			14.14	16.57	16.69
			15.68	13.61	18.35
			20.00	20.50	15.58
			31.84	31.53	25.71
			36.36	36.38	36.97

$$Y_{i_1, i_2, i_3} = (0.892 - 0.114 \sum_j S_j(i_j))^{-14.9}$$

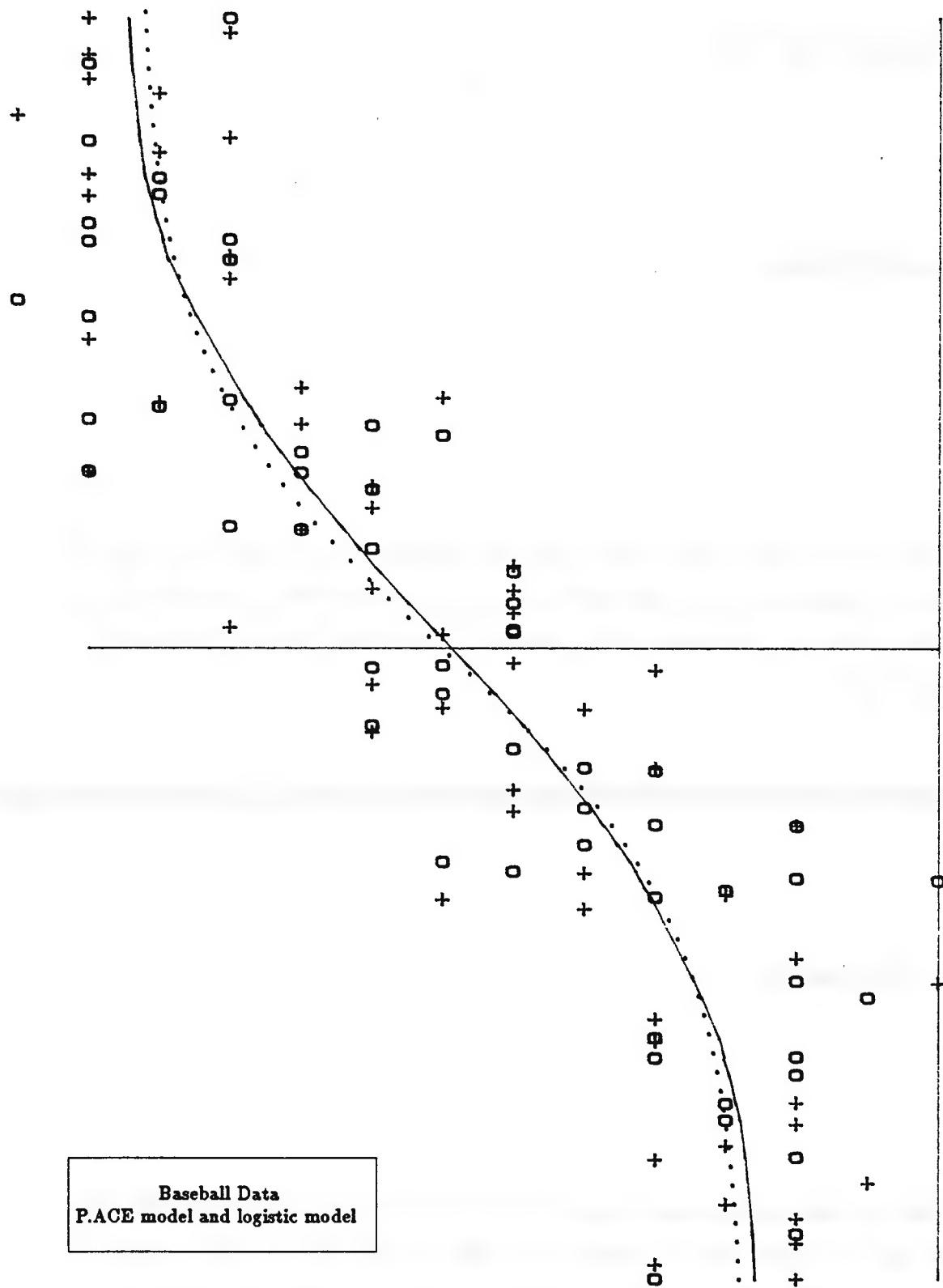


Baseball Data

WINS	L O S S E S								
	Cleveland	Boston	New York	Phil.	Detroit	St. Louis	Wash.	Chicago	
Cl.		11 10.2 11.2	10 12.0 11.5	16 13.1 13.4	13 13.2 14.1	14 15.0 15.0	16 15.1 15.1	16 15.4 15.2	
B.	10 11.4 10.8		14 12.5 11.3	12 13.5 13.2	15 13.8 14.0	15 15.2 15.0	15 15.3 15.1	14 15.4 15.2	
N.Y.	12 9.8 10.4	8 9.2 10.7		12 12.2 13.0	13 12.4 13.4	16 14.5 15.0	17 14.6 15.1	16 15.3 15.2	
P.	6 8.7 8.6	10 8.3 8.8	10 9.5 9.0		10 11.0 12.2	18 13.7 14.6	14 13.9 14.8	16 15.5 15.0	
D.	9 8.5 7.9	7 8.0 8.0	9 9.3 8.2	12 10.7 9.8		11 13.4 14.0	16 13.6 14.4	14 14.9 14.7	
S.L.	8 6.9 6.9	7 6.7 7.0	6 7.4 7.0	4 8.1 7.4	11 8.4 7.9		10 11.0 11.9	13 13.1 12.7	
W.	6 6.8 6.8	7 6.7 6.9	5 7.2 6.9	8 7.9 7.2	6 8.2 7.6	11 10.7 10.1		12 13.0 11.9	
Ch.	6 6.6 6.6	8 6.8 6.8	6 6.8 6.9	6 7.0 7.0	8 7.0 7.3	8 8.7 9.3	9 8.8 10.1		

observed value	
Pace est.	Logistic est.

Baseball Team	P.ACE scores	Logistic scores	games won
Cleveland	0.434	0.441	96
Boston	0.362	0.418	95
New York	0.240	0.383	95
Philadelphia	0.080	0.160	84
Detroit	0.051	0.028	78
St. Louis	-0.287	-0.370	59
Washington	-0.313	-0.463	55
Chicago	-0.566	-0.560	51



Chapter 4

Simulations

The algorithm has been demonstrated to perform reasonably well on a few “real” data examples. Its performance on sparse tables has yet to be examined. The intent of this chapter is to demonstrate the performance of the procedure under several different conditions using simulated data.

4.1. The Data.

All the simulated data arose from a $4 \times 4 \times 3 \times 2 \times 3$ contingency table. This table of 288 cells was considered an acceptable compromise between tables of larger size which were computationally very expensive to analyze and tables of smaller size which could be less informative. For all simulations the scores were fixed and were as follows:

predictor	scores
1)	0.10 - 0.0333333 - 0.10 0.0333333
2)	- 0.22 - 0.14 0.02 0.34
3)	0.20 0.20 - 0.40
4)	- 0.1597079 0.1597079
5)	- .1202921 - .08 0.2002921
	(4.1.1)

The scores for predictor one were chosen so that the inner two scores were separated by a relatively small distance while the outer two scores were at a much greater distance away. The intent was to see if given the relatively large separation of the outer scores and the relatively small separation of the inner scores if the P.ACE procedure could distinguish and separate the inner scores. The scores for predictor 2 are a scale and location shift of the sequence 8, 16, 32, 64. The geometrically increasing separation of consecutive scores offered another method of assessing the procedure's ability to distinguish within categories scores. Two of the scores of predictor 3 were chosen to be equal to see if the procedure would be able to detect this fact. The scores for predictor 4 were random. It was the author's intent to have them equal to $\pi/20$, although his skill in basic arithmetic prevented this from occurring. The scores for predictor 5 were what is known as a "kludge". The algorithm scales the scores so that the maximum of the absolute value of the sum of the

scores is 1. Hence it was desirable to have the maximum of the sum of scores to be 1 and for symmetry reasons to also have the minimum of the sum of scores to be -1. The two criteria were satisfied by setting the scores of predictor 5 to the values presented.

From the examples in chapter 3 three parametric curves seem to be appropriate; the linear curve, the exponential curve, and the logistic curve. The curves were defined on the range of the scores, i.e. $[-1, 1]$, and were standardized by scaling and shifting them to have the value of 0 at -1 and 1 at 1. The three curves used to generate the data were

$$1) \text{ linear } f(x) = \frac{x+1}{2}$$

$$2) \text{ exponential } f(x) = \frac{e^x - e^{-1}}{e - e^{-1}} \quad (4.1.2)$$

$$3) \text{ logistic } f(x) = \frac{e^{7.5x}}{1 + e^{7.5x}}$$

With the data generated, normal errors were then added. To test the procedure in the best circumstances a trial with no noise was run. In addition noise with twelve different values of standard deviations were also considered;

$$\sigma = .025, .05, .075, .10, .125, .15, .175, .20, .25, .30, .40, .50 \quad (4.1.3)$$

Several small values of σ were included to observe the influence of increasing the amount of noise in the data. The large values were included to observe the breakdown of the model. From the scaling of the curves used to generate the data a standard deviation of $\sigma = .125$, for example, corresponds to a noise to signal ratio of 12.5 %.

After generating the data eleven different proportions of data were removed from the table to observe the behavior of the algorithm in cases of missing data. The proportions chosen are listed below:

proportion removed	number removed	number left	
0	0	288	
5	14	274	
10	29	259	
15	43	245	
20	58	230	
25	72	216	
30	86	202	
40	115	173	
50	144	144	
60	173	115	
70	202	86	(4.1.4)

Two methods were chose to eliminate the data from the table.

- 1) random deletions — the data to be discarded as missing is chosen uniformly from the table. For example, if 29 observations are to be deleted, then with equal probability any of the $\binom{288}{29}$ sets of 29 numbers are chosen.
- 2) select deletions — this is a meager attempt to model some kind of dependence between the observations which are missing and their position in the table. The cells corresponding to the second category of the first predictor or the second category of the third predictor had a greater chance of being eliminated. More precisely, let $C_j(x)$ be the category of the j^{th} predictor corresponding the the cell entry x . Then if a total of x observations were to be removed from the table, $x/9$ would be removed from $\{x|C_0(x) = C_3(x) = 2\}$, $2x/9$ would be removed from $\{x|C_3(x) \neq C_0(x) = 2\}$, $x/9$ would be deleted from $\{x|C_0(x) \neq C_3(x) = 2\}$, and the remaining $x/3$ from $\{x|C_0(x) \neq 2, C_3(x) \neq 2\}$.

Thus, to summarize, a contingency table with 5 predictors X_1, \dots, X_5 was considered.

The response could be expressed as

$$Y_{i_1, \dots, i_5} = f\left(\sum_{j=1}^5 S_j(i_j)\right) + \epsilon \quad (4.1.5)$$

where the scores are given in (4.1.1) and the function is one of the functions in (4.1.2). The noise, ϵ , was normal with zero mean and standard deviation one of the values in (4.1.3). One of the several different proportions listed in (4.1.4) were removed by either random deletions or selected deletions.

For each combination 100 trials were run. In each case the curve generated by the procedure was evaluated at intervals of one tenth from -1 to 1 to obtain some measure of how close the curve found approximated the true curve. A problem arises when the maximum of the sum of true scores corresponding to the observations in the deleted table was less than one. If the procedure found the true curve, it would actually determine only a portion of the curve interior to the range $[-1, 1]$. Since the procedure used the scale convention that the maximum sum was one, the portion would be magnified by some constant and would have the affect of altering the apparent fit between the true and estimated curve. To lessen this, the procedure kept track of the "correct" scaling of the scores, extending by linear approximation those values which were outside the range of the approximating curve. Although this practice is not optimal, the two scales differed only rarely in tables which had many missing observations and the differences were usually quite small.

4.2. Results.

For each of the combinations listed above two measures of goodness of fit are calculated. The first measure is the standard deviation of the fit to the data. This is the square root of the average square deviations of the data to the value estimated by the P.ACE algorithm. The second is the standard deviation of the fit to the model. This is the square root of the average of the squared deviations of the true underlying curve to the curve estimated

by the P.ACE algorithm, evaluated at the points $\{-1.0, -0.9, \dots, 0.9, 1.0\}$. The results can be found in table Ia through table VIb at the end of this chapter.

The first observation concerns the deviation of the fit to the data. In all cases considered the error is relatively constant across the differing amounts of missing data and the type of deletion used. For the exponential and linear models, except for very low levels of noise to signal ratios the deviation of fit to the data is smaller than the standard deviation of the noise added to the data. This suggests that the procedure is over-fitting the data, although the difference between the two numbers is quite small in most cases and is unlikely to be very serious. The fit for the logistic model is somewhat worse, only a few cases having a smaller standard deviation to fit than the standard deviation of the noise added. This is most likely due to the fact that unlike the exponential and linear curves, the logistic curve is neither convex nor concave. The procedure appears to perform quite well in general when trying to fit straight lines and only slightly less well in fitting functions which deviate slightly from linear. The convexity of the exponential functions considered is relatively small, undoubtably accounting for some of the success of the procedure in that case. The logistic function, however, is highly non-linear and the procedure has difficulty in finding the exact curve.

The second observation concerns the standard deviation of the fit to the model. As would be expected, the fit degenerates as either the number of observations missing or the noise to signal ratio increases. What is important to note is that the deterioration of the fit is not very sensitive to the number of missing values. This indicates that the procedure may be of use when large amount of data is missing.

The ability of the procedure to locate the correct scores appeared to be very insensitive to either the type of deletions, the underlying model, the amount of noise or the amount of data, provided that noise to signal ratio was 30% or less. For this reason only one set of scores have been included. Page 63 contains the average scores obtained from the P.ACE procedure for the logistic model with 30% of the observations selectedly deleted and with a noise to signal ration of 25%. Included also are the true scores and the standard deviation of each score from its mean. As can be seen, there is close agreement between the estimated

scores and the true values, and the estimated scores tend to be rather stable. There seems to be a tendency to find scores having greater uniformity in between score spacings than exists in the original scores. The two inner scores of predictor 1 have been separated somewhat and the procedure has brought the scores of predictor 2 closer to a set of uniformly spaced scores. Nevertheless, these affects are rather slight.

To obtain some idea of how well the model estimated the the underlying curve two graphs were drawn for each of the underlying curves. To illustrate the ability of the procedure to find structure in sparse tables all the examples were chosen from tables missing 50 % of the observations. There seemed to be little dependence of the curves on the type of deletion present in the table. The tables with random deletions and 30 % noise to signal ratio and also the tables with select deletions with 15 % noise to signal ratio were chosen from each of the three different curves. The true value of the curve was drawn in circles at each of the points $\{-1.0, -0.9, \dots, 0.9, 1.0\}$. Also at these points the median, the quartiles, and the 5 and 95 percentiles were calculated from the 100 trials of the simulation and plotted. The medians were connected with a straight line, the quartiles with dotted lines and the 5 and 95 percentiles with dotted lines, giving some idea of how well the curves were approximated. The curves are rather self-explanatory. Just a few points should be mentioned. The flaring out at the ends is common in smoothers and is caused by the asymmetric position of the data in the window of the smoother. It should be noticed that interior to the interval $[-1, 1]$ the fit is rather good and the 5 and 95 percentiles are close to the true values. From the graphs it appears as if the procedure is somewhat median biased, although the curves corresponding to 15 % noise to signal ratio have very little bias in the interior of the range.

4.3. Conclusions.

The P.ACE algorithm presented offeres a method of fitting the model

$$Y_{i_1, \dots, i_j} = \theta_k \left(\sum_l S_l(i_l) \right) \quad (4.3.1)$$

to contingency tables. The model appears to be a reasonable model, corresponding to certain established parametric models such as the independent model or the logistic model. As shown in chapter 3, it also yields similar results to these classical models in those instances that the models describe the data well.

There appears to be several advantages of using the P.ACE algorithm and the model (4.3.1). The first advantage is in the generality of the model and the few assumptions made about the data. Because the model mimics the classical models well it can also be used as a means of model selection, and prove useful in exploratory analysis of contingency tables. In many cases the scores obtained by the procedure can help give some indication of the relative importance of each predictors, and can give some means of interpreting the effects certain predictors on the response.

The second advantage is the ability to perform well under conditions in which much of the data is missing. The method can provide estimates for the zero cells of a table while making a minimal set of assumptions and can help detect patterns in tables which may be difficult to observe otherwise.

There are also several disadvantages of the procedure. The procedure does not perform well when the contingency table is small or when the range of the response is large compared the the number of observations. The first problem is due to the fact that the procedure is estimating a large number of parameters and small tables do not provide enough information. The second problem is related to the smoothing algorithm incorporated into the procedure. The amount of possible curvature in the output of the smoother is a function of the bandwidth, which is bounded below by the number of data points. Thus small data sets can not produce a smooth curve with a large curvature. The yarn data of chapter 3 illustrates this point.

There is a definite lack of theory. It is straightforward to show that under general conditions an optimal set of scores and functions exist which minimize the mean square error

of the model (4.3.1). It is equally straight forward to construct examples of non-uniqueness of the solutions. There is also no proof of convergence of the algorithm, although in practice this had never presented any problem with the data used.

In spite of the shortcomings, however, the method presented seems to be a useful method in the exploratory analysis of contingency tables, and one worth pursuing.

4.4. Graphs and Charts.

This section contains the charts and graphs discussed in the earlier sections of this chapter.

Table Ia
 standard deviation of fit to data
 linear model with random deletions

		proportion of data missing										
		.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70
standard deviation	.000	.00	.01	.01	.01	.02	.02	.02	.03	.03	.04	.04
	.025	.02	.03	.03	.03	.03	.03	.03	.04	.04	.04	.05
	.050	.05	.05	.05	.05	.05	.05	.05	.05	.06	.06	.06
	.075	.07	.07	.07	.07	.07	.07	.08	.08	.08	.08	.08
	.100	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
	.125	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12
	.150	.15	.15	.15	.15	.15	.15	.15	.15	.15	.14	.14
	.175	.17	.17	.17	.17	.17	.17	.17	.17	.17	.17	.17
	.200	.19	.20	.19	.19	.20	.20	.19	.19	.19	.19	.19
	.250	.24	.24	.24	.24	.24	.24	.24	.24	.24	.23	.23
n	.300	.29	.29	.29	.29	.29	.29	.29	.29	.28	.28	.28
	.400	.39	.39	.39	.38	.39	.39	.38	.38	.38	.37	.36
	.500	.49	.48	.49	.49	.48	.48	.49	.48	.47	.47	.46

Table Ib
 standard deviation of fit to model
 linear model with random deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation of fit to model with random deletions	.000	.00	.02	.03	.04	.04	.05	.05	.06	.07	.07	.11
	.025	.01	.02	.03	.04	.04	.05	.05	.06	.08	.09	.11
	.050	.03	.03	.04	.04	.06	.06	.06	.06	.08	.10	.11
	.075	.04	.05	.05	.05	.06	.06	.07	.07	.09	.10	.13
	.100	.05	.06	.06	.07	.06	.07	.08	.08	.09	.10	.15
	.125	.07	.07	.08	.07	.09	.09	.10	.09	.12	.12	.14
	.150	.08	.08	.08	.09	.10	.09	.10	.10	.13	.13	.16
	.175	.09	.09	.10	.09	.10	.10	.12	.12	.12	.17	.17
	.200	.12	.11	.12	.12	.12	.11	.13	.14	.16	.19	.24
	.250	.13	.13	.14	.14	.14	.16	.16	.18	.21	.21	.26
standard deviation of fit to model with random deletions	.300	.15	.16	.17	.17	.18	.20	.19	.21	.25	.25	.31
	.400	.24	.25	.24	.27	.26	.29	.29	.35	.37	.42	.43
	.500	.30	.33	.33	.37	.40	.41	.39	.42	.46	.57	.56

Table IIa
standard deviation of fit to data
linear model with select deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
s t a n d a r d d e v i a t i o n	.000	.00	.01	.01	.01	.02	.02	.02	.03	.03	.04	.04
	.025	.02	.03	.03	.03	.03	.03	.03	.04	.04	.04	.05
	.050	.05	.05	.05	.05	.05	.05	.05	.06	.06	.06	.06
	.075	.07	.07	.07	.07	.07	.08	.08	.08	.08	.08	.08
	.100	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
	.125	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12
	.150	.15	.14	.14	.14	.15	.15	.15	.14	.14	.14	.14
	.175	.17	.17	.17	.17	.17	.17	.17	.17	.17	.17	.16
	.200	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	.250	.24	.24	.24	.24	.24	.24	.24	.24	.23	.23	.24
	.300	.29	.29	.29	.29	.29	.29	.29	.28	.28	.28	.27
	.400	.39	.39	.38	.38	.38	.39	.38	.38	.37	.37	.36
	.500	.48	.48	.48	.48	.48	.48	.47	.47	.46	.45	

Table IIb
 standard deviation of fit to model
 linear model with select deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation of fit to model linear model with select deletions	.000	.00	.01	.01	.02	.02	.02	.03	.04	.04	.06	.11
	.025	.01	.01	.02	.02	.02	.03	.03	.03	.04	.08	.08
	.050	.03	.03	.02	.03	.03	.03	.04	.04	.06	.07	.09
	.075	.04	.04	.04	.04	.04	.05	.05	.05	.08	.10	.15
	.100	.05	.07	.06	.06	.06	.06	.07	.07	.08	.11	.14
	.125	.07	.07	.06	.06	.08	.08	.08	.10	.11	.12	.17
	.150	.08	.08	.08	.09	.09	.10	.09	.10	.13	.13	.18
	.175	.09	.09	.09	.10	.10	.10	.11	.13	.16	.16	.22
	.200	.10	.11	.13	.12	.13	.14	.14	.15	.15	.19	.26
	.250	.14	.15	.16	.17	.18	.19	.20	.20	.25	.25	.34
standard deviation of fit to model linear model with select deletions	.300	.19	.20	.20	.22	.22	.24	.26	.28	.31	.36	.45
	.400	.24	.30	.29	.31	.34	.36	.36	.38	.39	.40	.54
	.500	.37	.38	.40	.39	.40	.45	.46	.47	.50	.54	.65

Table IIIa
standard deviation of fit to data
exponential model with random deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation	.000	.01	.02	.02	.02	.02	.02	.03	.03	.04	.04	
	.025	.03	.03	.03	.03	.03	.03	.04	.04	.04	.05	
	.050	.05	.05	.05	.05	.05	.05	.05	.06	.06	.06	
	.075	.07	.07	.07	.07	.08	.08	.08	.08	.08	.08	
	.100	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10	
	.125	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	
	.150	.15	.15	.15	.15	.15	.15	.15	.15	.14	.15	
	.175	.17	.17	.17	.17	.17	.17	.17	.17	.17	.16	
	.200	.20	.20	.19	.19	.19	.19	.19	.19	.19	.19	
	.250	.24	.25	.24	.24	.24	.24	.24	.24	.23	.23	
n	.300	.29	.29	.29	.29	.29	.29	.29	.28	.28	.27	
	.400	.39	.39	.39	.38	.39	.38	.38	.37	.37	.36	
	.500	.48	.48	.48	.48	.48	.48	.47	.47	.47	.44	

Table IIIb
 standard deviation of fit to model
 exponential model with random deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	=
standard deviation of fit to model with random deletions	.000	.02	.04	.04	.04	.04	.05	.06	.06	.06	.08	=
	.025	.04	.05	.05	.05	.05	.05	.05	.05	.07	.09	=
	.050	.06	.06	.06	.06	.06	.06	.06	.07	.09	.11	=
	.075	.07	.07	.06	.07	.08	.07	.08	.09	.09	.11	=
	.100	.08	.09	.09	.09	.09	.10	.10	.13	.14	.15	=
	.125	.09	.10	.10	.10	.11	.10	.11	.12	.14	.14	=
	.150	.11	.12	.12	.12	.13	.13	.14	.16	.16	.19	=
	.175	.13	.14	.15	.14	.15	.16	.16	.20	.20	.24	=
	.200	.14	.14	.17	.17	.17	.18	.17	.20	.21	.28	=
	.250	.18	.18	.23	.23	.25	.26	.26	.26	.27	.33	=
	.300	.24	.25	.25	.26	.25	.28	.29	.30	.31	.36	=
	.400	.32	.33	.32	.36	.35	.40	.43	.44	.48	.53	=
	.500	.49	.49	.48	.50	.50	.51	.54	.60	.61	.64	=

Table IVa
standard deviation of fit to data
exponential model with select deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation	.000	.01	.02	.02	.02	.02	.02	.03	.03	.04	.04	
	.025	.03	.03	.03	.03	.03	.03	.04	.04	.04	.05	
	.050	.05	.05	.05	.05	.05	.05	.05	.06	.06	.06	
	.075	.07	.07	.08	.08	.08	.08	.08	.08	.08	.08	
	.100	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10	
	.125	.12	.12	.12	.12	.12	.12	.12	.12	.12	.12	
	.150	.15	.15	.15	.15	.15	.15	.14	.14	.15	.14	
	.175	.17	.17	.17	.17	.17	.17	.17	.17	.17	.16	
	.200	.20	.20	.19	.19	.19	.19	.19	.19	.19	.19	
	.250	.24	.24	.24	.24	.24	.24	.24	.24	.24	.23	
n	.300	.29	.29	.29	.29	.29	.29	.29	.28	.28	.27	
	.400	.39	.39	.39	.39	.38	.38	.38	.37	.37	.36	
	.500	.48	.48	.48	.48	.48	.48	.47	.47	.46	.45	

Table IVb
 standard deviation of fit to model
 exponential model with select deletions

		proportion of data missing										
		.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70
standard deviation	.000	.04	.04	.04	.04	.04	.04	.04	.04	.05	.07	.11
	.025	.04	.04	.04	.04	.04	.04	.05	.05	.06	.07	.12
	.050	.05	.05	.05	.05	.05	.05	.06	.06	.07	.08	.11
	.075	.05	.06	.07	.07	.07	.07	.06	.07	.08	.09	.15
	.100	.07	.07	.08	.08	.07	.07	.08	.09	.11	.11	.13
	.125	.08	.08	.08	.08	.09	.09	.11	.10	.11	.12	.15
	.150	.09	.09	.10	.11	.12	.12	.12	.13	.14	.15	.22
	.175	.12	.12	.13	.14	.14	.14	.15	.16	.15	.18	.24
	.200	.12	.12	.14	.15	.16	.16	.17	.17	.18	.21	.22
	.250	.17	.18	.19	.19	.20	.20	.19	.22	.26	.26	.32
deviation	.300	.22	.23	.23	.24	.25	.26	.25	.29	.30	.33	.39
	.400	.26	.29	.32	.35	.37	.39	.39	.40	.41	.54	.57
	.500	.43	.45	.47	.48	.48	.47	.48	.49	.49	.53	.59

Table Va
standard deviation of fit to data
logistic model with random deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation of fit to data	.000	.04	.04	.04	.05	.05	.06	.06	.07	.09	.10	.12
	.025	.05	.05	.05	.05	.06	.06	.07	.08	.09	.10	.12
	.050	.06	.06	.07	.07	.07	.08	.09	.09	.11	.12	
	.075	.08	.08	.09	.09	.09	.10	.11	.11	.12	.13	
	.100	.11	.11	.11	.11	.11	.12	.12	.13	.14	.14	
	.125	.13	.13	.13	.13	.14	.14	.14	.15	.15	.16	
	.150	.15	.15	.15	.16	.16	.16	.16	.17	.17	.17	
	.175	.18	.18	.18	.18	.18	.18	.18	.19	.19	.19	
	.200	.20	.20	.20	.20	.20	.20	.21	.21	.21	.21	
	.250	.25	.25	.25	.25	.25	.25	.25	.25	.25	.25	
deletion	.300	.29	.30	.30	.30	.30	.30	.30	.30	.30	.29	
	.400	.39	.39	.39	.39	.39	.39	.39	.40	.39	.38	
	.500	.49	.49	.49	.49	.48	.48	.49	.48	.48	.47	

Table Vb
 standard deviation of fit to model
 logistic model with random deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation	.000	.00	.00	.00	.01	.01	.02	.02	.03	.05	.10	.15
	.025	.01	.01	.01	.02	.02	.02	.03	.05	.07	.10	.18
	.050	.02	.02	.03	.03	.04	.03	.04	.05	.11	.09	.17
	.075	.03	.03	.04	.04	.04	.04	.06	.06	.07	.11	.14
	.100	.04	.04	.05	.05	.05	.06	.05	.06	.09	.13	.15
	.125	.05	.05	.06	.06	.06	.06	.09	.09	.09	.13	.20
	.150	.07	.06	.07	.08	.08	.09	.09	.11	.11	.16	.19
	.175	.07	.09	.09	.10	.10	.08	.10	.10	.12	.15	.23
	.200	.08	.10	.11	.10	.11	.11	.11	.12	.16	.17	.23
	.250	.11	.11	.14	.12	.13	.13	.13	.16	.18	.19	.30
n	.300	.12	.15	.14	.16	.16	.17	.16	.19	.22	.20	.39
	.400	.21	.22	.20	.20	.23	.23	.21	.27	.31	.38	.46
	.500	.25	.26	.27	.30	.30	.33	.36	.35	.44	.52	.60

Table VIa
standard deviation of fit to data
logistic model with select deletions

proportion of data missing												
	.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	
standard deviation of fit to data	.000	.04	.04	.04	.05	.05	.06	.06	.07	.09	.10	.12
	.025	.05	.05	.00	.05	.06	.06	.07	.08	.09	.10	.11
	.050	.06	.06	.07	.07	.07	.08	.08	.09	.10	.11	.13
	.075	.08	.08	.09	.09	.09	.09	.10	.10	.11	.12	.13
	.100	.11	.11	.11	.11	.11	.12	.12	.13	.14	.15	
	.125	.13	.13	.13	.13	.14	.14	.14	.15	.15	.16	
	.150	.15	.15	.15	.16	.16	.16	.16	.17	.17	.18	
	.175	.18	.18	.18	.18	.18	.18	.18	.19	.19	.20	
	.200	.20	.20	.20	.20	.20	.20	.21	.21	.21	.21	
	.250	.25	.25	.25	.25	.25	.25	.25	.25	.25	.25	
on	.300	.30	.30	.30	.30	.30	.30	.30	.30	.30	.29	
	.400	.39	.39	.39	.39	.39	.39	.39	.39	.39	.38	
	.500	.49	.49	.49	.49	.49	.49	.49	.48	.48	.47	

Table VIb
 standard deviation of fit to model
 logistic model with select deletions

		proportion of data missing										
		.00	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70
standard deviation of fit to model with select deletions	.000	.00	.00	.00	.01	.01	.02	.02	.03	.05	.10	.15
	.025	.01	.01	1.00	.01	.02	.03	.03	.05	.05	.07	.16
	.050	.02	.02	.02	.03	.03	.04	.04	.05	.07	.09	.20
	.075	.03	.03	.03	.03	.05	.04	.05	.06	.08	.12	.20
	.100	.04	.04	.05	.04	.05	.06	.07	.07	.10	.16	.19
	.125	.05	.05	.06	.06	.07	.08	.07	.09	.12	.13	.23
	.150	.07	.06	.07	.08	.08	.08	.09	.11	.12	.15	.23
	.175	.07	.08	.08	.10	.10	.09	.09	.13	.14	.15	.26
	.200	.08	.09	.09	.10	.10	.11	.11	.14	.16	.18	.26
	.250	.11	.12	.12	.13	.14	.15	.14	.15	.19	.19	.28
standard deviation of fit to model with all deletions	.300	.13	.13	.13	.16	.17	.18	.18	.19	.21	.28	.33
	.400	.17	.20	.21	.20	.23	.24	.26	.26	.35	.37	.46
	.500	.28	.30	.29	.30	.35	.37	.38	.41	.41	.45	.55

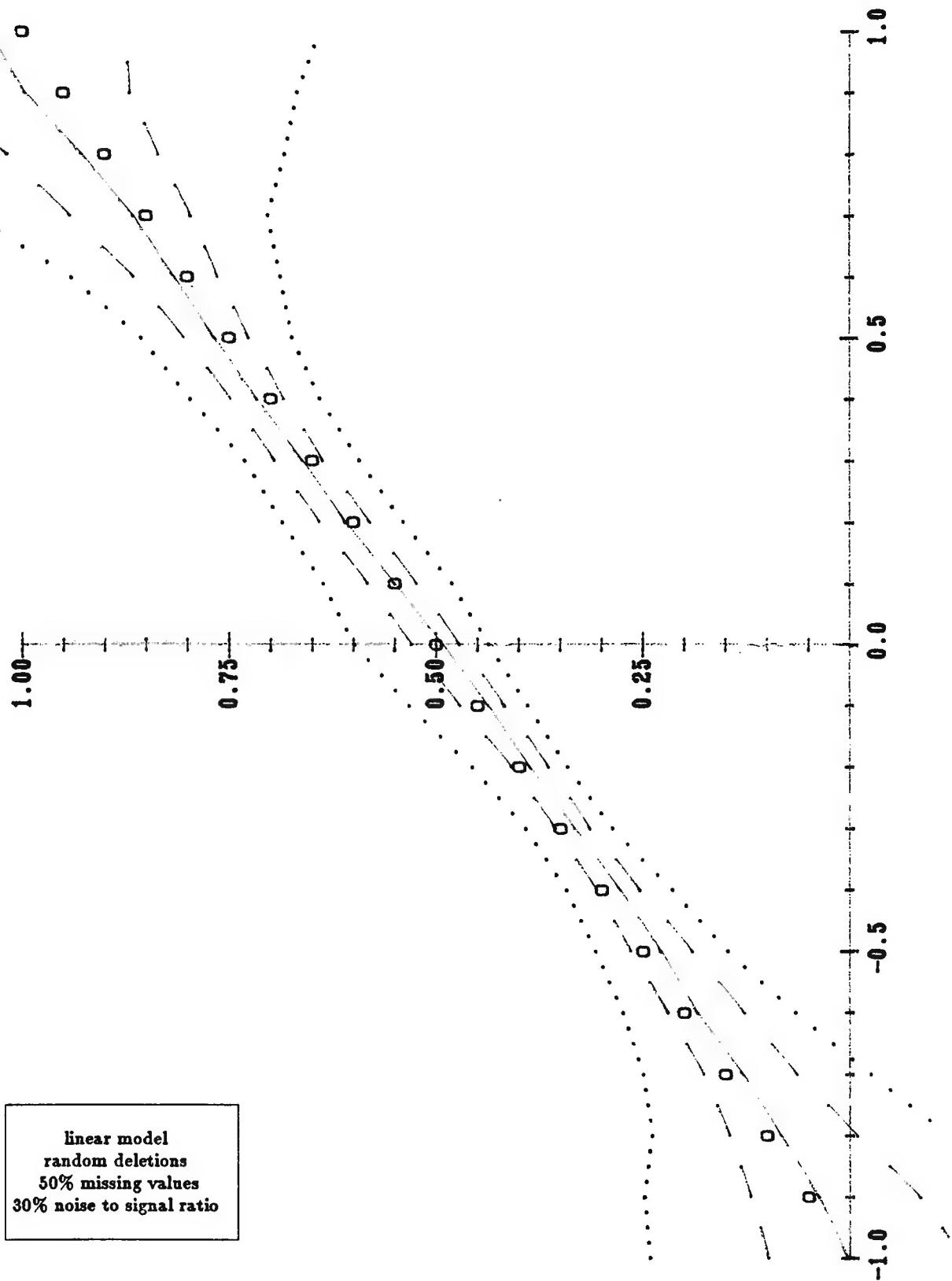
Scores found for Logistic Model

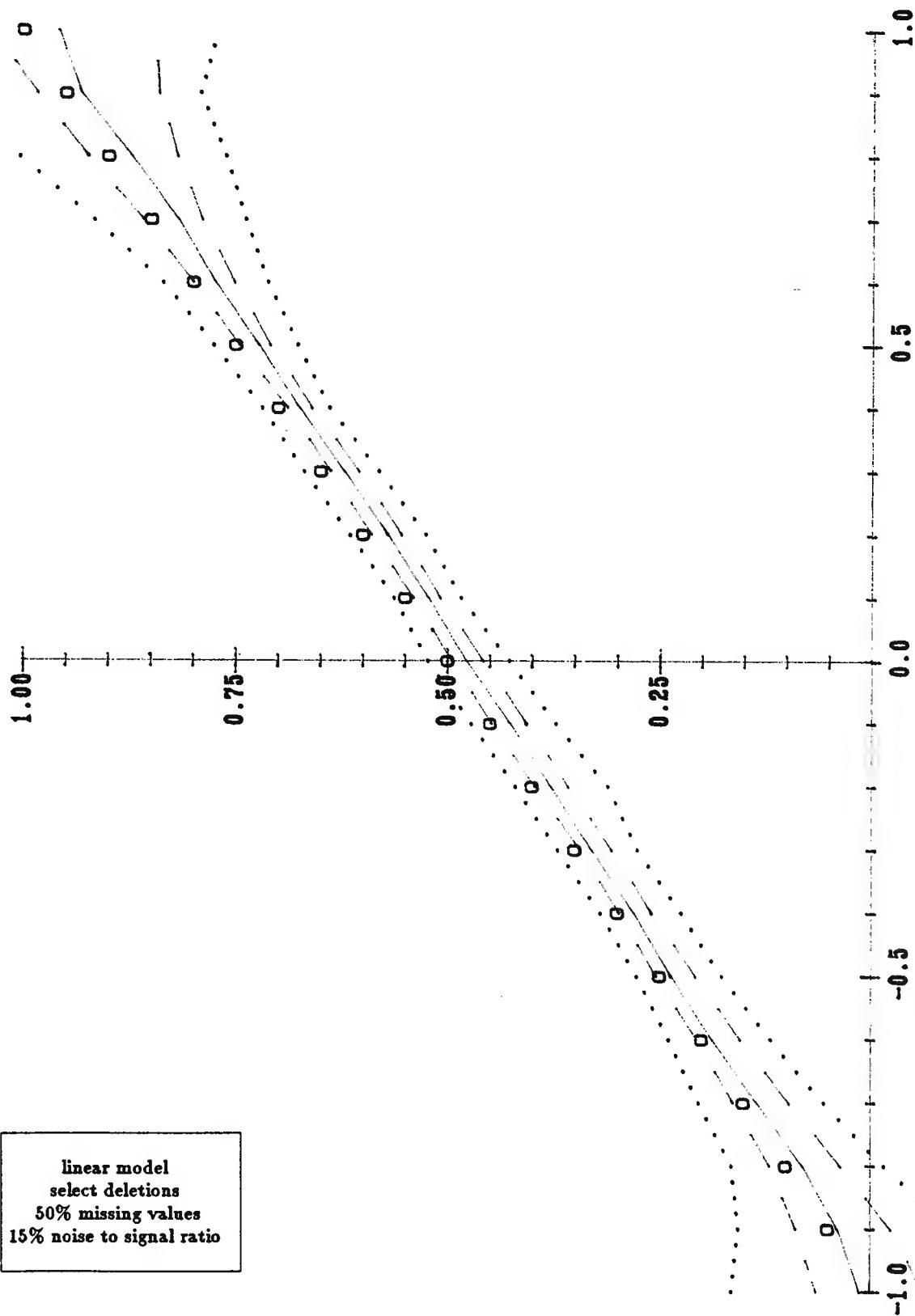
select deletions
 30% observations missing
 Standard Deviation = 0.25

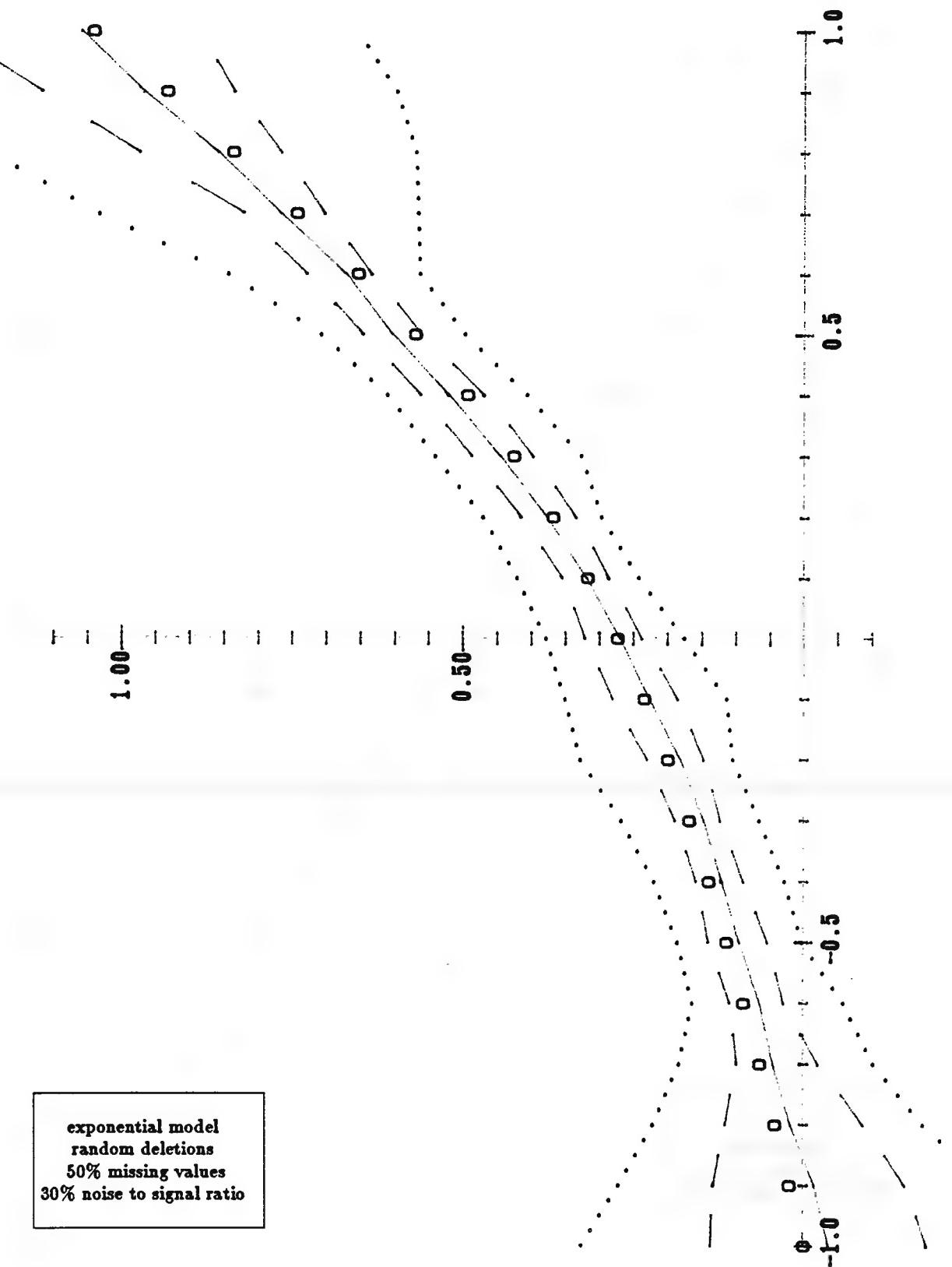
Predictor	True Scores*	Estimated Scores	standard deviation**
1)	0.100	0.104	3.71
	-0.033	-0.048	3.94
	-0.100	-0.100	3.52
	0.033	0.044	4.05
2)	-0.220	-0.206	3.55
	-0.140	-0.141	3.35
	0.020	0.014	4.32
	0.340	0.332	4.32
3)	0.200	0.188	3.00
	0.200	0.186	3.49
	-0.400	-0.373	4.56
4)	-0.160	-0.154	2.51
	0.160	0.154	2.51
5)	-0.120	-0.113	3.05
	-0.080	-0.077	3.38
	0.200	0.191	3.01

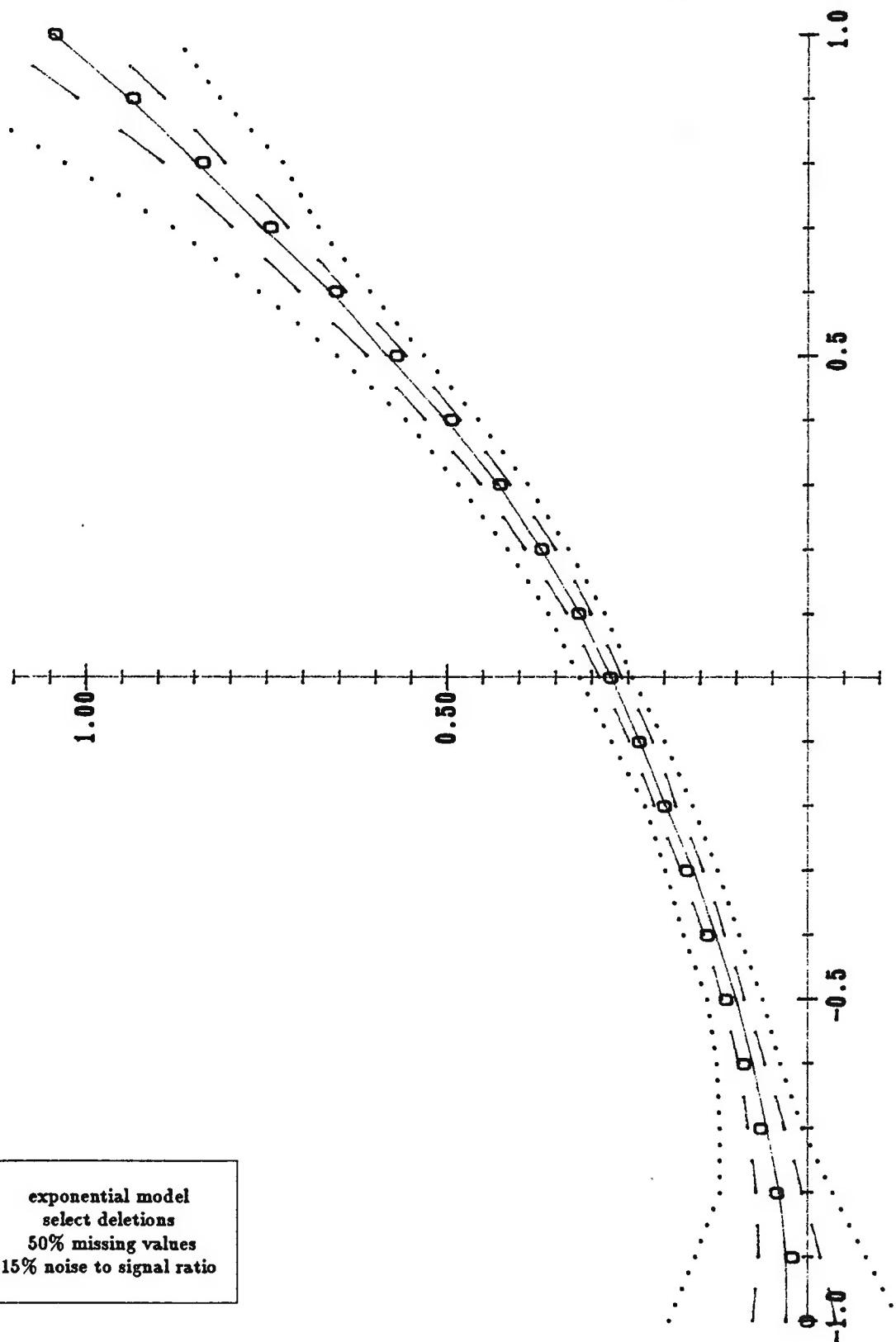
* Three significant digits given only

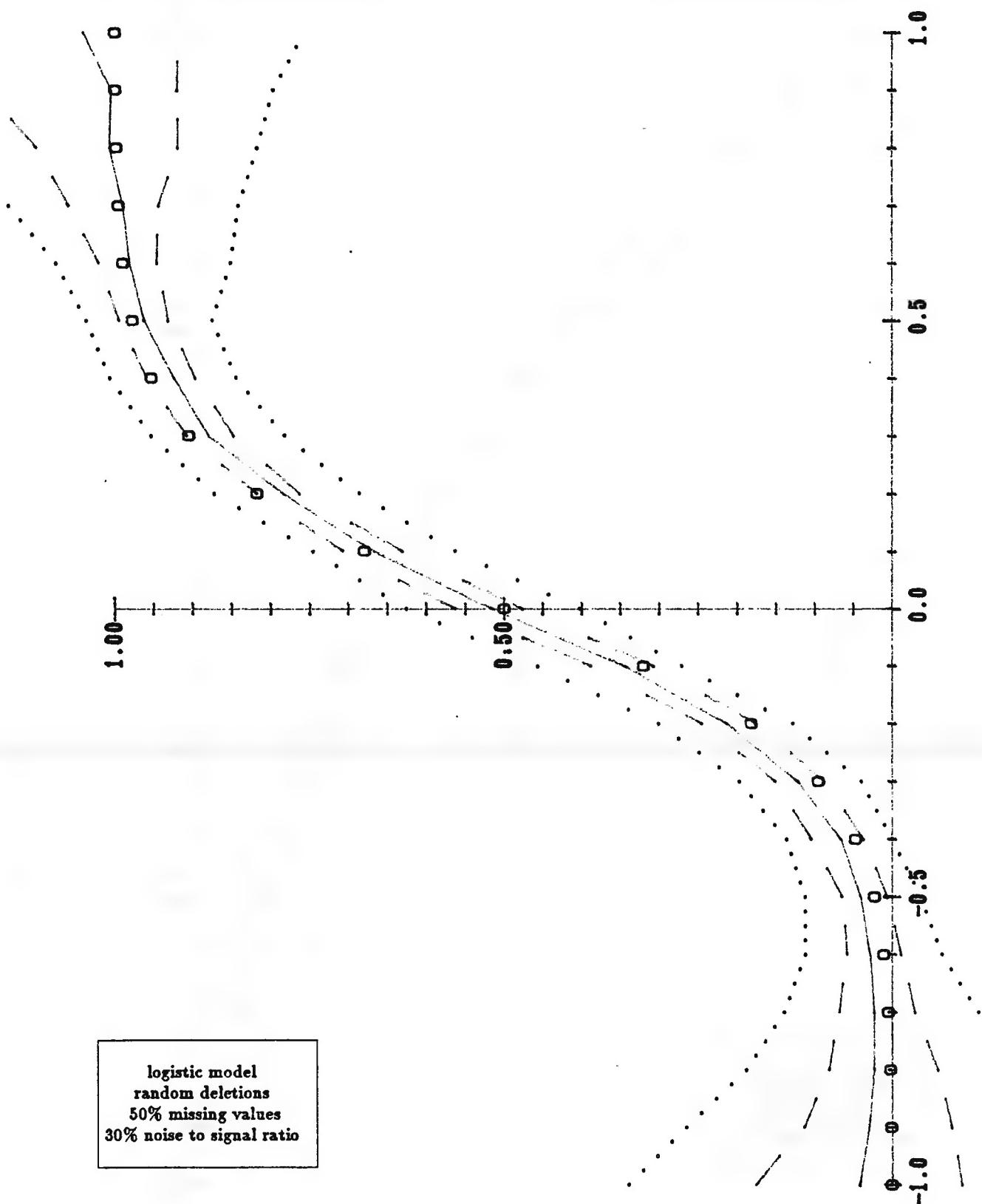
** Standard deviations expressed in terms of 10^{-2} .

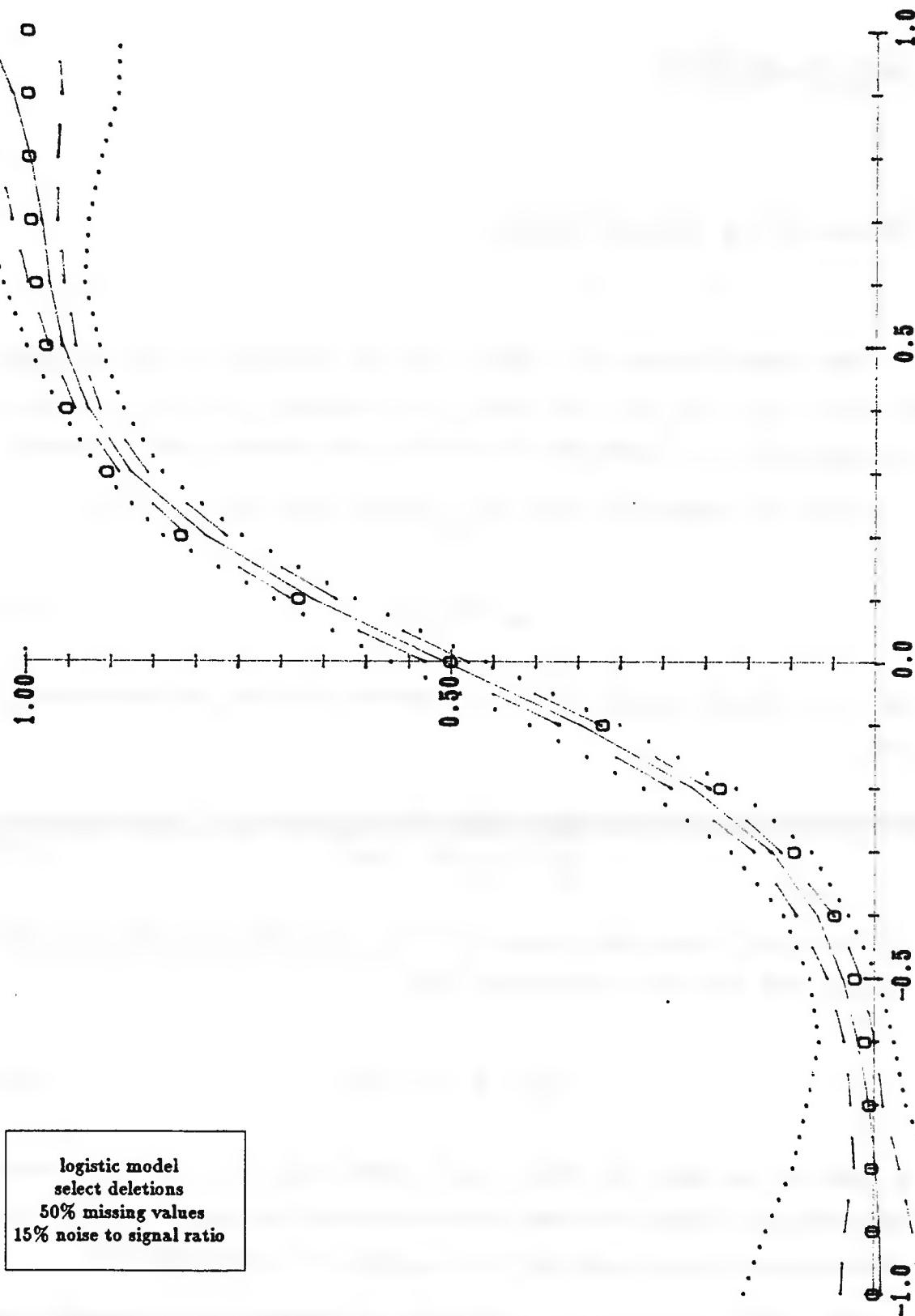












Appendix

Smoothing Algorithms

There is much literature on smoothing and the reader unfamiliar with these techniques is referred to [11], [18], [20]. What follows is an abbreviated discussion, concentrating primarily on the method employed in the P.ACE procedure demonstrated in this work.

Given a set of observations $\{(x_i, y_i)\}_{i=1}^n$, a possible summary of the data is

$$y_i = f(x_i) + r_i \quad (\text{A.1})$$

where $f(\cdot)$, called the smooth, satisfies some smoothness constraint and is chosen to minimize

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - f(x_i))^2. \quad (\text{A.2})$$

The model (A.1) is appropriate, in particular, if it is assumed that the underlying random variables which generated the observations satisfy

$$\mathbb{E}(Y | X = x) = g(x) \quad (\text{A.3})$$

in which case the smooth $f(\cdot)$ of (A.1) can be viewed as an estimate of the conditional expectation $g(x)$ in (A.3). A smoother is a procedure which has as input a set of bivariate observations and returns the smooth function satisfying the decomposition (A.1).

One method of estimating the smooth is by use of local linear least squares fits. Let $\{(X_i, Y_i)\}_{i=1}^n$ be a fixed set of n bivariate observations, assuming $X_j < X_m$ if $j < m$. At

each observation point X_j ; let $J_k(X_j)$ be the set of observations $\{(X_l, Y_l)\}_{l=\max(j-k-1,1)}^{\min(j+k,n)}$. Fix k . The local linear smoothing procedure defines the smooth at X_j to be the value of the least squares straight line of the points in $J_k(X_j)$ evaluated at the point X_j . Linear interpolation is used to extend the definition of the smooth to abscissa values other than those included in the observations.

The number $2k+1$ is known as the span and controls the variability of the output of the smoother. As the span increases the curvature as well as the variance of the estimated curve decreases while the bias increases. Thus when smoothing a scatterplot it is desirable to decrease the span in those regions in which the underlying curve has high curvature and the observations have small variance. Likewise, in those regions in which the underlying curve has small curvature and the observations have high variance a large span is desired.

One possible approach to a such a variable span smoother is to use a point-wise convex combination of different fixed span smoothers with weights depending on some goodness-of-fit measure. More precisely, let C_1, \dots, C_p be the smooth curves obtained by using p different spans of a local linear smoother. The smooth curve C_f generated by such a variable span smoothing procedure can be written as

$$C_f(X_j) = \sum_{k=1}^p w_k(X_j) C_k(X_j) \quad (\text{A.4})$$

where the weights $\{w_k(X_j)\}_k$ are, for each X_j , a partition of unity.

One simple method of defining these weights would be to let $w_k(X_j)$ be proportional to $(Y_j - C_k(X_j))^{-2}$. The difficulty with this approach is that $w_k(X_j)$ need not be close to $w_k(X_{j+1})$. This results in a curve $C_f(\cdot)$ which may have high frequency components. Consequently, some alternate definition is necessary which will insure that the weights vary "smoothly". One approach is to consider the cross-validated squared error-difference in some neighborhood of the point X_j . Fixing a number l , using this approach the weights can be expressed as

$$w_k(X_j) \propto \left[\sum_{m \in J_k(X_m)} (Y_m - C_k^*(X_m))^2 \right]^{-2} \quad (\text{A.5})$$

where $C_k^*(X_m)$ is the value at the point X_m of the least squares line fit to the points $J_k(X_m) - \{X_m\}$. The smallest span has on average the smallest difference between the corresponding curve and the observations. This is due to the fact that the observation is used to form its own prediction, and its relative influence increases as the span decreases. Thus there is a bias towards smaller spans due to overfitting. Since the observations contain noise, a small difference between the smooth and the observations does not imply a small difference between the smooth and the underlying curve $f(\cdot)$ in (A.1). Consequently a bias towards small spans does not guarantee a good overall fit. Cross-validation is a method of getting rid of the bias.

Such a smoothing technique was employed in the P.ACE procedure demonstrated in this work. A set of three fixed span local linear smoothers were implemented with spans 10%, 20% and 50% of the number of observations and the number l used to determine the neighborhood size of (A.5) was set to 10% of the number of observations.

There exist simple updating formulas for the local linear smoothers which allow an order n algorithm for the computation of the fixed span smoother. The calculations of the weights of (A.5) can be incorporated as part of the local linear smoother.

The P.ACE algorithm as implemented requires in addition to the smooth function some estimate of its derivative. The local linear smoother uses least square lines to define the smooth at a point X_j . Associated with the line is a slope m_j which can be used as an estimate of the slope of the curve at X_j . For each span at each observation X_j the slope of the least squares line is stored. A convex combination of the slopes at each observation point using the weights defined by (A.5) is then formed. Although this would seem to yield a reasonable estimate of the derivative practice has shown that the resulting curve tends to contain high frequency components. Consequently, to dampen these components a local linear smoother with span 40% of the observations is applied to the convex combination and the resulting curve is taken as the estimate of the derivative.

Bibliography

- [1] Aickin, M. (1983) *Linear Statistical Analysis of Discrete Data*, New York: Wiley & Sons
- [2] Barlow, Bartholemew, Bremmer and Brunk (1972) *Statistical Inference Under Order Restrictions*, New York: Wiley & Sons
- [3] Bartlett M. S. (1935) "Contingency Table Interactions", *Journal of the Royal Statistical Society Supplement*, **2**, 248 — 252
- [4] Berkson, J. (1955) "Maximum likelihood and minimum chi-square estimates of the logistic function", *Journal of the Applied Statistical Association*, **50**, 130 — 162
- [5] Bhapker, V. P. and Koch, G. G. (1968) "Hypothesis of 'No Interaction' in Multidimensional Contingency Tables", *Technometrics*, **10**, 107 — 123
- [6] Birch, Y. M. M. (1963) "Maximum Likelihood in Three-Way Contingency Tables", *Journal of the Royal Statistical Society, series B*, **25**, 220 — 233
- [7] Bishop, Y. M. M. (1969) "Full contingency tables, logit, and split contingency tables.", *Biometrics*, **25**, 383 — 400
- [8] Bishop, Y. M. M. (1971) "Effects of Collapsing Multidimensional Contingency Tables", *Biometrics*, **27**, 545 — 562
- [9] Box, G.E.P., and Cox,D.R. (1964) "An analysis of transformations", *Journal of the Royal Statistical Society, series B*, **26**, 211 — 252
- [10] Breiman, L. and Friedman, J. (1982) "Estimating optimal Transformation for multiple regression and correlation", *Department of Statistics, Stanford University*, Tech. Report ORION 06
- [11] Cleveland, W. S. (1979) "Robust locally weighted regression and smoothing scatterplots", *Journal of the American Statistical Association*, **74**, 828 — 836

- [12] **Craig, C. C.** (1953) "Combination of Neighboring Cells in Contingency Tables", *Journal of the American Statistical Association*, **48**, 104 — 112
- [13] **Deming, W. E. and Stephan, F. F.** (1940) "On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Totals are Known", *Annals of Mathematical Statistics*, **11**, 427 — 444
- [14] **Fienberg, S. E.** (1970) "Quasi-independence and maximum likelihood estimation in incomplete contingency tables", *Journal of the American Statistical Association*, **65**, 1610 — 1616
- [15] **Fienberg, S. E.** (1980) *The Analysis of Cross-Classified Categorical Data (second addition)*, Cambridge Massachusetts: The MIT Press
- [16] **Fienberg, S. E. and Holland, P. W.** (1973) "Simultaneous estimation of multinomial cell probabilities", *Journal of the American Statistical Association*, **68**, 683 — 691
- [17] **Fisher, R.** (1938) *Statistical methods for Research Workers (tenth addition)*, Edinburgh: Oliver and Boyd
- [18] **Friedman, J.** (1984) "A Variable Span Smoother", *Department of Statistics, Stanford University, Tech. Report 5*
- [19] **Friedman, J. and Owen, A.** "in progress"
- [20] **Friedman, J. and Tibshirani, R.** (1983) "The monotone smoothing of scatterplots", *Department of Statistics, Stanford University, Tech. Report ORION 23*
- [21] **Goodman, L. A.** (1968) "The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries", *Journal of the American Statistical Association*, **63**, 1091 — 1131
- [22] **Goodman, Leo A.** (1971) "The analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classification", *Technometrics*, **13**, 33 — 61

- [23] Kendall, M. G. and Stuart, A (1977) *The Advanced Theory of Statistics* (fourth edition), London: C. Griffin
- [24] Lancaster H. D. (1951) "Complex contingency tables treated by the partition of chi-square", *Journal of the Royal Statistical Society, series B*, **13**, 242 — 249
- [25] Pearson, K (1900) "On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophical Magazine*, **50**, 157 — 175
- [26] Quetelet, A. (1827) "Recherches sur la Population, les Naissances, etc., dans le Royaume des Pays-Bas", *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelle*, **4**, 117 — 174
- [27] Rényi, A. (1959) "On measures of dependence", *Acta Math. Acad. Sci. Hung.*, **10**, 441 — 451
- [28] Roy, S. N. and Kastenbaum, M. A. (1956) "On the hypothesis of no 'interaction' in a multiway contingency table", *Annals of Mathematical Statistics*, **27**, 749 — 757
- [29] Simonoff, J. S. (1983) "A penalty function approach to smoothing large sparse contingency tables", *Annals of Statistics*, **11**, 208 — 218
- [30] United Nations (1982) *Demographic Yearbook - 1980*, New York: United Nations Publication
- [31] Young, F. W. (1981) "Quantitative analysis of qualitative data", *Psychometrika*, **46**, 357 — 388
- [32] Yule, G. U. (1900) "On the association of attributes in statistics: with illustrations from the material of childhood society, &c.", *Philosophical Transaction of the Royal Society, Series A*, **194**, 257 — 311